Intensity-dependent spatial summation

Tom N. Cornsweet and John I. Yellott, Jr.

Cognitive Science Group, University of California, Irvine, California 92717

Received November 13, 1984; accepted April 16, 1985

Psychophysical evidence indicates that, in the human retina, the size of the spatial-summation area decreases as illuminance increases. Such a relationship would be beneficial for the detection of spatial contrast in the presence of photon noise. We analyze an image-processing mechanism in which the area of a strictly positive point-spread function varies inversely with local illuminance while its volume remains constant. In addition to its expected effect of improving spatial resolution as illuminance increases, this mechanism also yields center–surround antagonism and all other manifestations of bandpass filtering and accounts for Ricco's law and Weber's law—including the failures of both laws as a function of test conditions. The relationship between this mechanism and lateral inhibition is analyzed.

1. INTRODUCTION

Many psychophysical and physiological experiments can be interpreted as showing that light falling upon any one point of the retina creates an excitatory effect at neighboring points and that this lateral excitation combines additively with the direct excitation produced by light itself.[1] Psychophysical evidence also indicates that the extent of lateral excitation—the size of the spatial-summation area—increases as retinal illuminance decreases.[2,3]

One obvious and undesirable consequence of spatial summation is, in effect, to blur the neural image, and so it is natural to look for compensatory benefits of the process. A plausible suggestion is that intensity-dependent spatial summation is an adaptive response to the intrinsic noisiness of light. If the effective flux density in an image is I (absorbed photons/unit time)/unit area, then both the mean and the variance of the actual quantum catch per unit time over an area A equal IA. This statistical relationship imposes a fundamental constraint on spatial contrast detection.

Suppose that a change in illuminance from I to $I + cI$ is to be detected with an error rate of the order of 0.001 and that the visual system is a perfect detector limited only by quantal fluctuations. Then the effects of the incident quanta must be summed over an area A large enough that[4]

$$IA > 10/c^2.$$

Thus, to detect a 100% contrast change ($c = 1$) lasting one time unit, IA, the total number of quanta whose effects are summed during one time unit must be greater than 10. To detect a contrast of 1% requires that $IA > 100{,}000$.

Individual human photoreceptors collect quanta over areas of the order of 10^{-5} mm^2 and integrate their quantum catch over temporal durations of the order of 0.1 sec. Taking absolute threshold to be 100 quanta/0.1 sec at the cornea, spread over a retinal area of the order of 10^{-3} mm^2, and assuming that 10% of corneal quanta are effectively absorbed by photopigment, I at the absolute threshold of human vision is of the order of 10^4 (quanta/0.1 sec)/mm^2. Therefore the value of IA for an individual receptor at absolute threshold is only about 1/100th of that needed to detect 100% contrast reliably and about 10^{-6} that needed to detect 1% contrast. Thus, if no spatial summation occurred, a 100% contrast could be detected only when retinal illuminance reached 100 times the absolute threshold level (a statement that is self-contradictory, since the absolute threshold is a contrast detection), and 1% contrast could not be detected until the illuminance was of the order of 10^6 times absolute threshold (that is, around 1 cd/m^2). Spatial summation can thus be seen as a device for pooling the retinal quantum catch over areas larger than a single receptor, allowing reliable contrast detection at scotopic and mesopic light levels. And the fact that the summation area becomes smaller as illuminance increases can be interpreted as an adjustment that tends to keep the summation area A as small as possible at each light level I, subject to a requirement of the form $IA > 10/c^2$, thereby minimizing needless reductions in spatial resolution.

This noise-compensation interpretation of spatial summation is well known, especially through the seminal work of Rose.[5] However, it does not seem to be widely recognized that an adaptive spatial-summation mechanism can automatically create effects resembling a number of well-known visual phenomena not generally associated with photon noise, including edge enhancement (Mach bands) and other bandpass-filter effects usually attributed to lateral inhibition. We have analyzed a model visual system based on the following assumption: Each point in the retinal image gives rise to a nonnegative point-spread function whose height is directly proportional to image intensity at that point and whose volume remains constant—so that the area covered by the point spread varies inversely with local image intensity. The output image is the sum of the point-spread functions generated around each input point. We refer to this operation as "intensity-dependent spatial summation."

This simple operation proves to have a surprising number of immediate consequences that resemble important features of human vision. It creates Mach bands at edges, sombrero-shaped impulse responses, and a low-frequency falloff in the spatial contrast-sensitivity function. [In fact, when the point-spread function is Gaussian, it yields the same contrast-sensitivity function (CSF) as a linear lateral inhibitory model whose point-spread function is the negative Laplacian of a Gaussian, as in the theory of Marr and Hildreth.[6]] In addition, the same assumption implies Weber's law (including its failures as a function of light intensity and target size) and Ricco's law (including the fact that the area of perfect spatial

0740-3232/85/101769-18$02.00

summation shrinks as the background light level increases) and causes visual acuity (the high-frequency cutoff of the CSF) to increase as the square root of mean luminance.[7] These consequences are robust under changes in the exact shape of the point-spread function (i.e., square, triangular, Gaussian, etc.) and depend only on the fundamental assumption that the area under that function is inversely proportional to local image intensity.

Finally, it is noteworthy that this spatial-summation mechanism mimics not only the main effects usually attributed to lateral inhibition, such as Mach bands, but also the apparent dependence of lateral inhibition itself on the mean luminance level. For example, the response to small spots has a distinct sombrero form only when the spot is superimposed upon a relatively high-intensity background. When background intensity is low the "negative" brim of the sombrero becomes vanishingly small, as though lateral inhibition failed at low light levels—a result that has been reported for retinal ganglion cells[8,9] and that is also found in psychophysical measurements of spatial contrast sensitivity.[10,11] Here, however, there is never any inhibition—all the model's consequences are due to changes in the width of a nonnegative point-spread function. A similar realistic dependence on background intensity also appears in the model's response to other stimulus configurations commonly used in psychophysical experiments. For example, the background intensity level beyond which detectability of a target obeys Weber's law shifts upward as the area of the target decreases.[12]

Organization

In this paper we describe the basic mathematical properties of image processing by intensity-dependent spatial summation. Our purpose is to introduce a theoretical tool that may prove useful in visual system modeling and also in image-processing technology. In Section 2 we define the simplest intensity-dependent spatial summation (IDS) operator and derive some general results used repeatedly later on. In Sections 3 and 4 we describe the effects of applying this IDS operator to images commonly used in psychophysical measurements of spatial contrast sensitivity, such as edges, spots, and gratings. By and large, these effects are qualitatively in agreement with the results of psychophysical experiments, but we point out some significant differences and comment on their implications. We also note similarities between the consequences of IDS processing and physiological results frequently cited as demonstrations of lateral inhibition in the retina. In Section 5 we discuss the relationship between IDS operators and linear operators commonly employed in visual theory and the potential value of IDS operators in artificial image processing. In Section 5 we also describe a generalized IDS operator that retains the basic properties of the model introduced in Section 2 and allows a better fit to psychophysical data.

Although IDS is in a sense motivated by photon-noise considerations, this paper focuses on its consequences for deterministic input images, for which analytic results can be obtained relatively easily. That is not so for Poisson noisy images, which apparently must be approached by simulation methods and properly form the subject for another paper.

2. THE INTENSITY-DEPENDENT SPATIAL-SUMMATION MODEL

Figure 1 illustrates the basic ideas of the IDS model. A two-dimensional input image (here, a sharp edge) is recorded by an array of photoreceptors, and they feed into a summation network that performs the IDS operation. That operation consists of two stages. First, each receptor gives rise to a nonnegative point-spread function whose center height is directly proportional to the intensity of the input image at that receptor and whose volume is constant—so that its area (that is, the volume divided by the center height) is inversely proportional to the input intensity. Second, these point-spread functions are added together to create the output image. That image is then read out over an array of output channels—one for each receptor location.

In this section we define the general class of IDS operators, give an example based on Gaussian point-spread functions, and derive some useful technical results. In Section 3 we work out the response properties of IDS models for a variety of one-dimensional input images, and in Section 4 we do the same for two-dimensional inputs. Whenever possible we derive the general properties that characterize the model's responses independent of the exact shape of the point-spread function. Then in every case we give the specific form of the response for the special case of a Gaussian point-spread function and illustrate the profile of that response graphically.

For mathematical convenience, our analytic treatment assumes that the photoreceptors are infinitely small relative to the size of the input and the output images. That is, we deal with the continuous case, in the same spirit as theories that model retinal processing by a convolution of continuous retinal images with continuous impulse responses. This continuous approximation to the discrete nature of actual retinas and man-made image processors provides realistic results up to input image intensity levels that would cause the point-spread function to become narrower than a single receptor or a single pixel.

Notation and Assumptions

$I(x, y)$ denotes the input image intensity at point (x, y); $\mathbf{O}[I(x, y)](p, q)$ denotes the output image intensity at point (p, q) when the input image is $I(x, y)$. (p and q refer, respectively, to the x and the y coordinates in the output image plane.) When the input image is obvious, we occasionally denote the output image simply as $\mathbf{O}(p, q)$.

The basic idea of the model is that each input point (x, y) contributes a nonnegative point-spread value to every output point (p, q), the size of the contribution depending on the input intensity value $I(x, y)$ and the distance from (x, y) to (p, q). Thus we need to specify a spread function of the general form $S\{(x, y), (p, q), I)\}$ that gives the contribution from (x, y) to (p, q) when the input intensity at (x, y) is I. We assume first that

(1) S is nonnegative.

(2) S is spatially homogeneous and circularly symmetric. (That is, S can be written as a function of two real variables in the form $S\{[(x - p)^2 + (y - q)^2], I\}$.)

Next we formalize the fundamental assumption that the

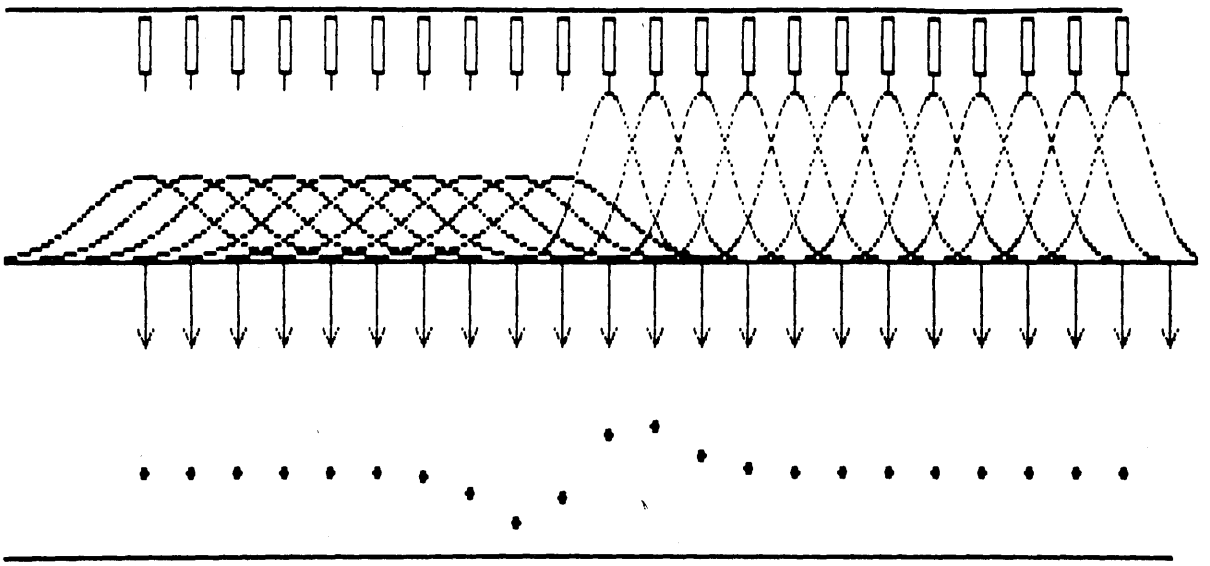

Fig. 1. Schematic diagram of the IDS model. From top to bottom: input image profile (here, a sharp edge); photoreceptors; photoreceptor point-spread functions (for the Gaussian case of the model); output channels (arrows); output image profile (dots).

area covered by the point-spread function around each input point varies inversely with the input intensity at that point. To accomplish this we assume that the center height $S(0, I)$ is directly proportional to the input intensity I, while the volume under S remains constant for all nonzero values of I:

(3) $S\{[(x-p)^2+(y-q)^2], I\} = I \times S\{I \times [(x-p)^2+(y-q)^2], 1\}$.

For any spread function S, integrating the right-hand side of assumption (3) over p, q yields a constant value V_s that is independent of I, while the height at the center [i.e., $S(0, I)$] equals $I \times S(0, 1)$. So the equivalent area under the point-spread function around any input point (volume divided by center height) is $1/I$ times the constant $V_s/S(0, 1)$. The choice of the volume constant V_s is arbitrary; it simply sets the value of the model's baseline response to uniform-field inputs, as is shown below in Theorem 1. We take this to be unity.

(4) The integral of $S\{[(x-p)^2+(y-q)^2], I\}$ over the p, q plane equals 1.0.

Given assumption (4), the remaining constant $1/S[0, 1]$ equals the equivalent area of the point-spread function when the input intensity $I = 1$. This parameter determines the numerical values of the point-spread areas for all input intensities and needs to be chosen appropriately to fit the model to psychophysical data. We make no specific assumption here about its value since that will depend on the units used to measure retinal area and light intensity.

In view of assumption (3), the point spread S is really a function of a single variable, so we can suppress the redundant intensity variable and express the fundamental assumption of the model as follows.

The point spread from input point (x, y) *to output point* (p, q) *is*

$$I(x, y) \times S\{I(x, y) \times [(x-p)^2+(y-q)^2]\},$$

where $I(x, y)$ *is the input image intensity at* (x, y) *and* S *is a nonnegative real function for which*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(p^2+q^2)\mathrm{d}p\mathrm{d}q = 1.$$

Different cases of the model can then be created by different choices of the basic spread function S, i.e., S may be Gaussian (as in the example below), square, exponential, etc. However, as we shall see, the exact choice makes little difference.

Note that the functional form of the spread function remains constant as $I(x, y)$ varies. For any input intensity I the point spread takes the form $I \times S(Ir^2)$, where S is a fixed function and r is distance from the input point. Thus the effect of the input intensity at each point is simply to rescale the spread function, leaving its basic form unchanged. As will be seen below in Theorem 3 and subsequently, this is an important feature of the model.

Finally, we assume that the output image is the sum of the point-spread functions:

$$(5)\quad \mathbf{O}[I(x, y)](p, q) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x, y) \times S\{I(x, y) \times [(x-p)^2+(y-q)^2]\}\mathrm{d}x\mathrm{d}y.$$

Assumption (5) entirely captures the notion of an IDS operator.

Example: The Gaussian Case

Suppose that S is the Gaussian function

$$S(x^2+y^2) = (1/2\pi) \times \exp[(-1/2) \times (x^2+y^2)]$$

corresponding to the joint probability density function (pdf) of two independent normal random variables, each with mean zero and variance one. Then the point spread around an input point (x, y) with intensity $I(x, y)$ is

$$[I(x, y)/2\pi] \times \exp\{(-1/2) \times I(x, y)[(x-p)^2+(y-q)^2]\},$$

i.e., a bivariate normal density function, centered at that point, corresponding to the joint pdf of two normal random variables, each having variance $1/I$. Figure 2 illustrates this point-spread function for several values of I. We use this example

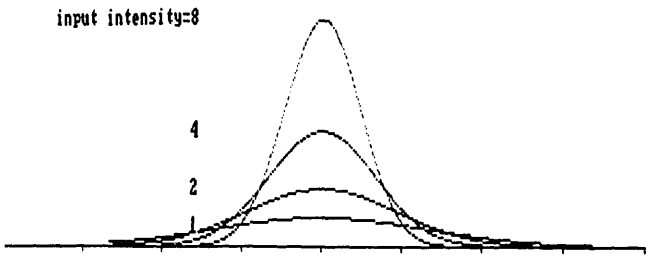

Fig. 2. Point-spread functions of the Gaussian case of the IDS model shown for four input intensities.

throughout to illustrate the model. Mathematically it is uniquely convenient because the Gaussian is the only circularly symmetric function that is also separable. However, as was noted earlier, the effects of IDS are largely independent of the exact shape of the spread function. To demonstrate this, our theorems are proved for arbitrary spread functions that satisfy assumptions (1)–(4).

This Gaussian version of the model has a point spread whose effective width is $6/\sqrt{I}$. Assuming a photoreceptor width of $1/150$-deg visual angle (2 μm), the point spread would shrink to a single receptor when I becomes greater than 800,000. We have confined our examples to I values less than 10,000 to keep the results of our continuous analysis realistic. In the figures below, the spatial units are degrees. The graphs show output image profiles over a retinal distance of ±2 deg, plotted at 150 points/deg.

Preliminary Results

An easy way to see that the model is nonlinear is to note the following.

Theorem 1

The output to any nonzero uniform field is the uniform field 1.0. [That is, when $I(x, y) \equiv I > 0$, $\mathbf{O}(p, q) \equiv 1$.]

Proof

Put $I(x, y) = I$ in assumption (5) and make the change of variable $u = (x - p)\sqrt{I}, v = (y - q)\sqrt{I}$. (Note: The output to a zero-intensity uniform field is again a zero-intensity field. Thus it might seem that there is a discontinuity in the uniform-field response. In practice this is not so, because any real input image is limited in spatial extent, whereas Theorem 1 assumes a truly infinite uniform field. For uniform-field inputs of any finite size, the response can be made as near zero as desired by making the input intensity sufficiently low.)

The physical meaning of Theorem 1 can be understood in the following way. Because the volume under the spread function at each point is constant and independent of the input intensity, the total output of the system is independent of its input—the effect of any input image is not to change the total amount of output but only to change its spatial distribution. Since a spatially uniform input image must generate a uniform output image, it follows that the output amplitudes corresponding to all uniform input images must be identical.

The next theorem simply documents a property built in by assumption (2): The IDS model is invariant under translations and rotations.

Theorem 2

If the input image is translated or rotated by any amount, the output image is unchanged except for translation or rotation by the same amount.

Proof

For translation: To represent a translation of the output to image $I(x, y)$ {i.e., $\mathbf{O}[I(x, y)](p - j, q - k)$} put $p = p - j, q = q - k$ in assumption (5) and make the change of variable $u = x + j, v = y + k$. This yields

$$\iint I(u - j, v - k) \times S\{I(u - j, v - k) \times [(u - p)^2 + (v - q)^2]\}du\,dv,$$

which is the output for the translated input image $I(x - j, y - k)$. (Note: To simplify notation we omit the limits of integration in this expression and those below. Unless otherwise noted, these can always be assumed to be the entire plane.)

For rotation: To represent a rotation of the output to $I(x, y)$ by a counterclockwise angle θ we substitute $p \cos\theta + q \sin\theta$ for p and $q \cos\theta - p \sin\theta$ for q in assumption (5) and make the change of variable $x = u \cos\theta + v \sin\theta, y = v \cos\theta - u \sin\theta$. Expanding the squared terms, we get

$$\iint I(u \cos\theta + v \sin\theta, v \cos\theta - u \sin\theta) \times S\{I(u \cos\theta + v \sin\theta, v \cos\theta - u \sin\theta) \times [(u - p)^2 + (v - q)^2]\}du\,dv,$$

which is the output for the rotated input image $I(x \cos\theta + y \sin\theta, y \cos\theta - x \sin\theta)$.

The final theorem of this section describes the effect of multiplying all the input image intensities by a common factor—i.e., the effect of changing the input image from $I(x, y)$ to $c \times I(x, y)$, as would happen with the retinal image of a real scene if the illumination falling upon that scene changed. This simple theorem is really the mathematical heart of the model: From it we can prove that Weber's law holds at edges, that Ricco's law holds for spots on a dark background, and that visual acuity increases in proportion to the square root of the mean luminance level—all regardless of the specific form of the point-spread function.

Theorem 3 (Scaling Theorem)

For every positive constant c and every input image $I(x, y)$

$$\mathbf{O}[cI(x, y)](p, q) = \mathbf{O}[I(x/\sqrt{c}, y/\sqrt{c})](p\sqrt{c}, q\sqrt{c}). \quad (1)$$

In words, this means that the effect of multiplying all the intensities in the input image by a constant c is the same as first expanding the original image spatially by a factor $\sqrt{c}$ along both axes, then applying the summation operator in assumption (5) to that image, and finally shrinking the output image back to the original size. Thus, for example, each spatial frequency f in the image $cI(x, y)$ is treated like frequency $f/\sqrt{c}$ in the image $I(x, y)$.

Proof

The right-hand side of Eq. (1) is

$$\iint I(x/\sqrt{c}, y/\sqrt{c})S\{I(x/\sqrt{c}, y/\sqrt{c}) \times [(x - p\sqrt{c})^2 + (y - q\sqrt{c})^2]\}dx\,dy.$$

Making the change of variable $u = x/\sqrt{c}$, $v = y/\sqrt{c}$, we obtain

$$\iint cI(u,v) \times S\{cI(u,v) \times [(u-p)^2 + (v-q)^2]\}dudv,$$

which is the left-hand side of Eq. (1).

3. RESPONSES TO ONE-DIMENSIONAL PATTERNS: EDGES, BARS, AND GRATINGS

Suppose that the input image is intrinsically one dimensional, i.e., $I(x,y) = I(x)$. (Because of Theorem 2, it is sufficient to consider only vertical one-dimensional inputs.) Making this substitution in assumption (5), we have

$$\mathbf{O}[I(x)](p,q) = \int \sqrt{I(x)} \int \sqrt{I(x)} S\{[(x-p)\sqrt{I(x)}]^2 + [(y-q)\sqrt{I(x)}]^2\}dydx.$$

Now in the inner integral (over y) we make the change of variable $v = (y-q)\sqrt{I(x)}$ to obtain

$$\mathbf{O}[I(x)](p) = \int_{-\infty}^{\infty} \sqrt{I(x)}\$\{(x-p)\sqrt{I(x)}\}dx, \qquad (2)$$

where \$ is the line-spread function corresponding to S, given by

$$\$(x) = \int_{-\infty}^{\infty} S(x^2+y^2)dy. \qquad (3)$$

It is easily seen that \$ is always nonnegative, symmetric about the origin, and integrates to 1. In the Gaussian example we have

$$\$(x) = \int_{-\infty}^{\infty} (1/2\pi)\exp[(-1/2)(x^2+y^2)]dy = (1/\sqrt{2\pi})\exp[(-1/2)x^2],$$

so the line spread around a line with intensity I is a normal pdf centered on the line, with variance $1/I$. Thus for the Gaussian case the response to one-dimensional patterns is given by

$$\mathbf{O}[I(x)](p) = \int_{-\infty}^{\infty} [\sqrt{I(x)}/\sqrt{2\pi}] \times \exp[(-1/2)I(x)(x-p)^2]dx. \qquad (4)$$

Step Response

Suppose that $I(x)$ is an edge of the form $I(x) = I$ for $x \le 0$, $I(x) = I + D$ for $x > 0$ (that is, a step). Then, for the Gaussian case, Eq. (4) yields the response

$$\mathbf{O}(p) = N[x(I+D)^{1/2}] + N[-x\sqrt{I}], \qquad (5)$$

where N is the cumulative normal distribution function:

$$N[z] = \int_{-\infty}^{z} (1/\sqrt{2\pi})\exp[(-1/2)x^2]dx.$$

Figure 3 shows the Gaussian-model step response [i.e., Eq. (5)] for a number of edges. These edges differ in illuminance (that is, I), but the ratio of the lighter to the darker side is the same for all—i.e., the ratio $(I+D)/I$, and consequently the Weber fraction D/I, is a constant. (Here $D/I = 10$.) It can be seen that the response displays Mach bands symmetrically located on either side of the step. At the step itself the response is always 1.0

To understand intuitively how Mach bands can be created

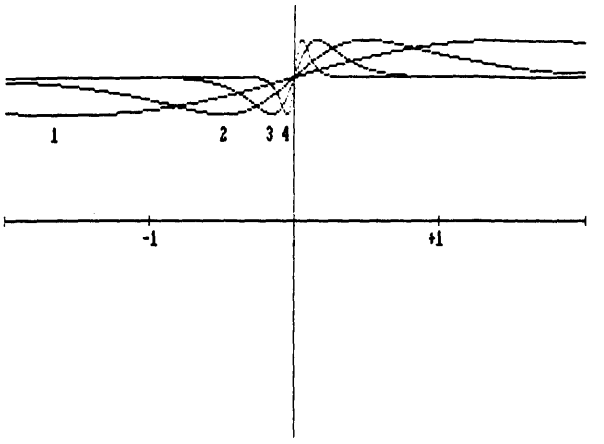

Fig. 3. Edge-response profiles. The input image was a step at zero from intensity I to $I + D$. $I/D = 10$ in all cases. Curve 1, $I = 0.1$; curve 2, $I = 1$; curve 3, $I = 10$; curve 4, $I = 100$.

by a purely positive point-spread mechanism (i.e., without lateral inhibition) it may be helpful to reexamine Fig. 1, bearing in mind that the output at each point is the sum of the spread functions above that point. As the edge is approached from the left (i.e., from the low-intensity side), the output decreases below the baseline level because there is less spread excitation coming from receptors on the right-hand side of the edge, which have narrower spread functions. Conversely, as the edge is approached from the high-intensity side, the output rises above the baseline level because of the extra excitation contributed by receptors on the low-intensity side, which have wider spread functions.

A second important feature of the response profiles in Fig. 3 is that the effect of increasing I is to move the peak and the trough of the Mach bands closer to the edge itself, but their amplitudes remain the same. This is a consequence of the fact that the input edges here all have the same Weber fraction D/I. Analysis of Eq. (5) shows that the peak of the positive-going Mach band occurs at $P\max = [(1/D)\log(1 + D/I)]^{1/2}$, and its value there is

$$\mathbf{O}(P\max) = N\{[(1 + I/D)\log(1 + D/I)]^{1/2}\} + N\{-[(I/D)\log(1 + D/I)]^{1/2}\},$$

which is a function only of the ratio D/I. The trough of the negative-going Mach band occurs at $P\min = -P\max$, and the output value there is $1 - [\mathbf{O}(P\max) - 1]$ (i.e., the peak is as far above the baseline response 1.0 as the trough is below it.) Thus the peak and trough values of the step response depend only on the Weber fraction D/I. Assuming a downstream detector mechanism that registers a perturbation in an otherwise uniform field when the output value at any point differs from the baseline 1.0 response by more than some threshold value, it follows that the Gaussian version of the model implies Weber's law for edge detection.

This result is not unique to the Gaussian case of the IDS model. Instead it holds for all cases [i.e., for all choices of the point-spread function S that satisfy assumptions (1)–(4)]. The following theorem shows why.

Theorem 4

Suppose that $I(x,y)$ is a straight edge separating a uniform field of intensity I from a field of intensity $I + wI$. Then the

maximum and minimum values of the output to $I(x, y)$ are independent of I and depend only on the Weber fraction w.

Proof

Because of Theorem 2 it is sufficient to consider only vertical edges of the form $I(x, y) = I(x) = I$ (for $x > 0$); $= I + wI$ (for $x \geq 0$). Suppose that $V(x)$ is a vertical edge image defined by $V(x) = 1$ for $x < 0$; $= 1 + w$ for $x \geq 0$. Assume that the maximum value of the output $\mathbf{O}[V(x)](p)$ occurs at $p = P\text{max}$ and that the minimum value occurs at $p = P\text{min}$. Let $I(x) = I$ for $x < 0$ and $I + wI$ for $x \geq 0$. Then $I(x) = I \times V(x)$, and so from Theorem 3 we have

$$\mathbf{O}[I(x)](p) = \mathbf{O}[I \times V(x)](p) = \mathbf{O}[V(x/\sqrt{I})](p\sqrt{I}) = \mathbf{O}[V(x)](p\sqrt{I}).$$

[The last equality holds because here $V(x/\sqrt{I}) = V(x)$.] The maximum value of the last expression in this line occurs at $p\sqrt{I} = P\text{max}$ and its minimum at $p\sqrt{I} = P\text{min}$, and so the maximum (minimum) output to $I(x)$ occurs at $p = P\text{max}/\sqrt{I}$ ($p = P\text{min}/\sqrt{I}$) and has the same value there that the output to $V(x)$ has at $P\text{max}$ ($P\text{min}$).

Two other features of the Gaussian-case step response can also be shown to be common to all IDS models: the fact that the output value at the step itself is always 1.0 and the fact that the locations of the peak and trough of the response move closer to the step as the baseline input-intensity level I increases. (The latter is true under the conditions that prevailed in Fig. 3, i.e., the edge separates fields of intensities I and $I + D$, and the Weber fraction D/I remains constant while I changes.)

To prove the first point, suppose that the input image is a vertical edge of the form $I(x, y) = I$ for $x \leq 0$ and $I + D$ for $x > 0$. We are concerned with the value of the output image $\mathbf{O}(p, q)$ along the vertical axis $p = 0$, and since it is sufficient to consider only a single point, we pick the origin [i.e., the point $(p, q) = (0, 0)$]. Then, from assumption (5), the output for an arbitrary spread function S is

$$\mathbf{O}(0, 0) = \int_{-\infty}^{\infty}\int_{-\infty}^{0} I \times S\{I[x^2 + y^2]\}dxdy + \int_{-\infty}^{\infty}\int_{0}^{\infty} (I + D) \times S\{(I + D)[x^2 + y^2]\}dxdy.$$

We know that $I \times S\{I[x^2 + y^2]\}$ is a circularly symmetric function whose integral over the entire x, y plane is 1.0, and the first integral in the expression above integrates this function over the half-plane $x \leq 0$, so its value must be 0.5. The same argument applies to the second integral, and consequently the entire expression equals 1.0.

Now to show that the distance from the edge to the locations of the maximum and minimum output values decreases as I increases, we can use the fact, shown in the proof of Theorem 4, that if $P\text{max}$ is the location of the maximum when the edge separates fields of intensities 1 and $1 + w$, then the maximum occurs at $p = P\text{max}/\sqrt{I}$ when the fields are I and $I(1 + w)$. So the distance between the location of the maximum and the edge itself varies inversely with $\sqrt{I}$. The same result for the minimum follows from the same argument.

The main result of this section is that for all IDS models, the step response always satisfies Weber's law. The same is also true of the response to bars and spots with sharp edges, provided that they are large—meaning large enough that there is no interaction between the responses to their two opposite edges. The next subsection should clarify this point.

Bar Response

Again, because of Theorem 2, it is sufficient to consider only vertical bars. Suppose that $I(x, y) = I(x) = I$ (a positive constant) for $|x| > W/2$; $I(x) = I + D$ for $|x| \leq W/2$ (so the input is a bar of width W and intensity D superimposed upon a uniform field of intensity I). Then the output for the Gaussian model is

$$\mathbf{O}(p) = N[\sqrt{I} \times (p - W/2)] + N[-\sqrt{I} \times (p + W/2)] + N[(I + D)^{1/2} \times (W/2 - p)] - N[-(I + D)^{1/2} \times (W/2 + p)]. \quad (6)$$

The form of the bar response depends on the bar width W and the background intensity I. Figure 4 illustrates the width effect: A narrow bar on a fairly intense background produces a response whose profile is sombrero shaped, quite like the line-spread function of a linear lateral-inhibitory model based on a difference of Gaussians or the negative Laplacian of a Gaussian. A wide bar of the same intensity on the same background produces Mach bands at both edges, and inside the edges the response returns to the baseline response value, just as would be expected from a linear model whose modulation transfer function (MTF) vanishes at the origin. The peak and trough amplitudes of the Mach bands in this case

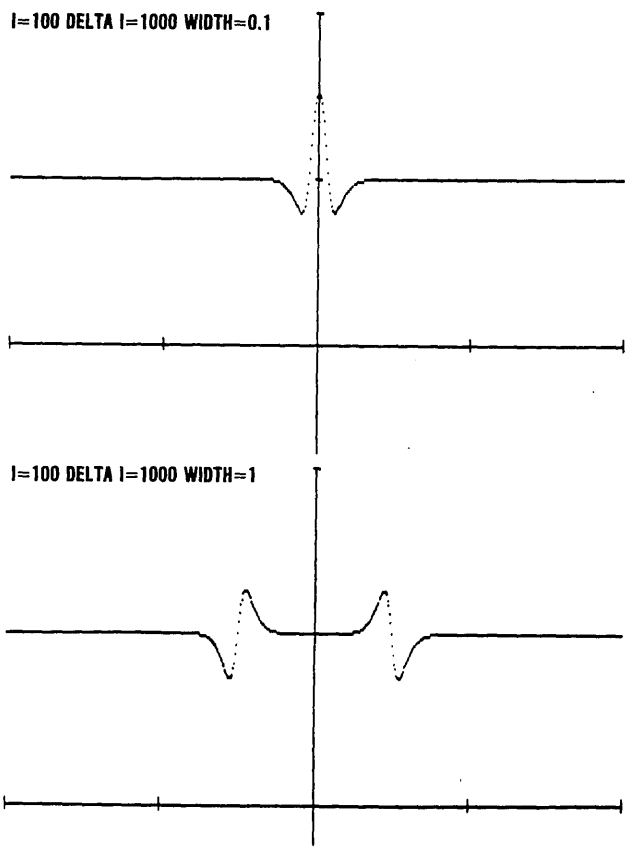

Fig. 4. Bar-response profiles for a narrow bar (top) and a wide bar (bottom) on a high-intensity background. Background intensity, I; bar intensity, $I + \Delta I$. Bar widths are as indicated in the figure. Tick marks on the abscissa indicate a width of 1.0.

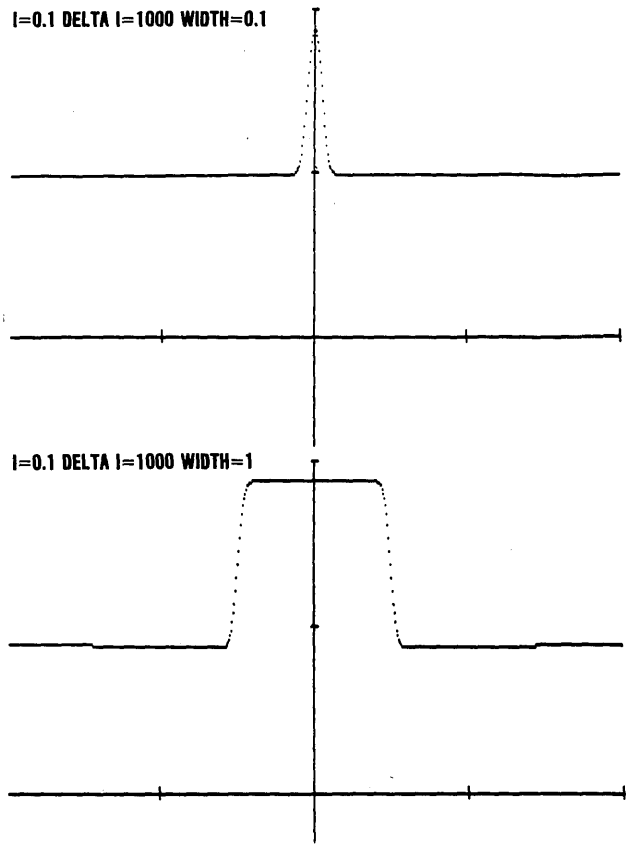

Fig. 5. Response profiles for the same bars as in Fig. 4 when the background has low intensity.

depend only on the Weber fraction D/I, so the detectability of wide bars should obey Weber's law.

Figure 5 shows output profiles for the same narrow and wide bars, but now superimposed upon a low-intensity background. The top panel illustrates how in this case the inhibitory lobes of the response to the narrow bar disappear (or, more precisely, become so broad and attenuated as to be unnoticable), and only the central excitatory portion of the response is evident. Thus "lateral inhibition" apparently fails when the background intensity is low—the receptive fields lose what appear to be their antagonistic surrounds and seem now to consist only of positive centers.

The bottom panel of Fig. 5 shows that the response to a wide bar also changes dramatically when the background intensity changes from high to low. Instead of a pair of narrow positive and negative Mach bands at both edges separated by an internal region of baseline-level output, the response now appears to be uniformly high within the bar, and outside each edge there is a broad negative Mach band. (If this bar were made much wider, the response inside its edges would eventually return to the baseline value, so that each edge would exhibit both positive and negative Mach bands. In general, the response profile for any target depends on its size relative to the background illuminance level.) From the standpoint of a classical receptive-field analysis it might appear that large receptive-field units retain their antagonistic surrounds at low light levels, whereas small units lose them—perhaps because of insufficient quantum catches in the regions feeding the smaller units. In an IDS system all these effects are due to intensity-dependent changes in the area of positive spatial summation.

Sinusoidal Grating Response

Suppose that the input is a sinusoidal grating of the form $I(x, y) = I(x) = L(1 + k \cos 2\pi fx)$: L is the mean intensity level, k is the grating contrast, and f is its spatial frequency. Because our operator is nonlinear we know that it must produce some harmonic distortion. Figure 6 shows the Gaussian-model response to high- (90%) and low- (20%) contrast sinusoidal grating inputs. At high contrast levels distortion is apparent: It takes the form of a spurious second harmonic that creates noticeable dimples at the peaks of the response. For low contrast levels, however, the output closely approximates a pure sine wave. Appendix A shows that for the Gaussian model the output to a low-contrast sinusoidal grating of the form $I(x) = 1 + k \cos 2\pi fx$ is approximately

$$\mathbf{O}(p) = 1 + \{2\pi^2 f^2 \exp[-2\pi^2 f^2]\} k \cos 2\pi fp. \quad (7)$$

The approximation given by Eq. (7) is obtained by solving Eq. (4) for $I(x) = 1 + k \cos 2\pi fx$ under the assumption that $k^2 = 0$. Consequently it is quite accurate for input contrasts on the order of 10% or less.

For low-contrast sinusoidal grating inputs, then, the outputs of the model are effectively sinusoidal, and it makes sense to speak of its MTF—i.e., the ratio of output contrast to input contrast as a function of input frequency. Let $G(f, L)$ denote the MTF for mean input level L. Equation (7) shows that

$$G(f, 1) = 2\pi^2 f^2 \exp(-2\pi^2 f^2). \quad (8)$$

To obtain the general form of the MTF we use the scaling theorem:

$$\begin{aligned}&\mathbf{O}[L(1 + k \cos 2\pi fx)](p)\\ &\quad = \mathbf{O}[1 + k \cos 2\pi fx/\sqrt{L}](p\sqrt{L})\\ &\quad = 1 + \{2\pi^2(f/\sqrt{L})^2 \exp[-2\pi^2(f/\sqrt{L})^2]\} k \cos 2\pi fp.\end{aligned}$$

So the MTF is

$$G(f, L) = 2\pi^2(f/\sqrt{L})^2 \exp[-2\pi^2(f/\sqrt{L})^2]. \quad (9)$$

Figure 7 shows this MTF for a range of mean intensity levels, plotted in the conventional way on log–log coordinates. In this plot the MTF shifts bodily to the right as L increases: Its peak (the best frequency) occurs at $f = (1/\pi\sqrt{2}) \times \sqrt{L}$, and visual acuity (defined as the highest frequency for which the MTF exceeds any fixed threshold) increases directly as $\sqrt{L}$.

Both the bandpass characteristics of the MTF and its bodily shift with changes in mean luminance (when the frequency

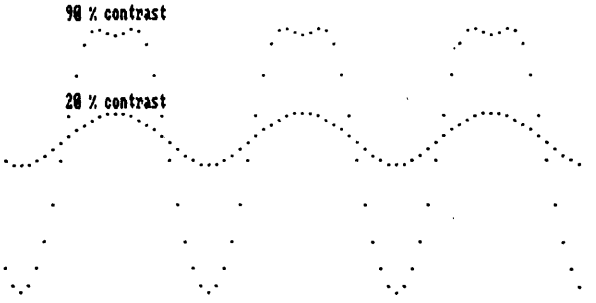

Fig. 6. Sinusoidal grating response profiles for high-contrast (90%) and low-contrast (20%) gratings.

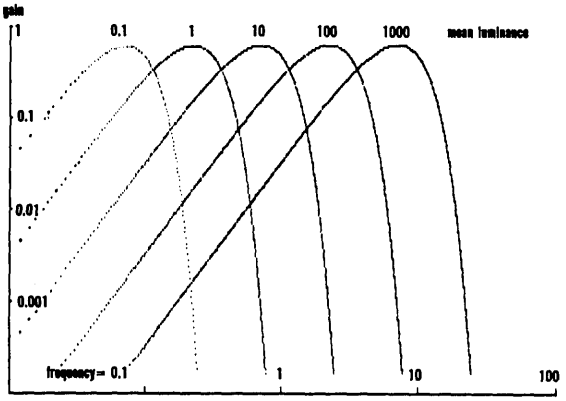

Fig. 7. MTF's of the Gaussian IDS model for various input mean luminance levels.

axis is logarithmic) are general properties of IDS models, independent of the exact form of the spread function S. Bandpass properties follow from the fact that very low frequencies will act like uniform fields and be driven to the baseline-response level, and very high frequencies will be attenuated by the basic point-spread operation. Bodily shifts with mean luminance follow from the scaling theorem, as is shown by the following.

Theorem 5

Suppose that for some range of contrast values the output to a sinusoidal input of the form $I(x) = 1 + k \cos 2\pi fx$ is another sinusoid of the form $\mathbf{O}(p) = 1 + G(f)\, k \cos 2\pi fp$. Then for any mean intensity level L the output to the sinusoidal input $L(1 + k \cos 2\pi fx)$ is $1 + G(f/\sqrt{L})k \cos 2\pi fp$. [In other words, the MTF at mean intensity L is $G(f/\sqrt{L})$.]

Proof

From Theorem 3

$$\begin{aligned}\mathbf{O}[L(1 + k \cos 2\pi fx)](p) &= \mathbf{O}[1 + k \cos 2\pi f(x/\sqrt{L})](p\sqrt{L}) \\ &= 1 + G(f/\sqrt{L})k \cos 2\pi(f/\sqrt{L})(p\sqrt{L}) \\ &= 1 + G(f/\sqrt{L})k \cos 2\pi fp.\end{aligned}$$

Consequently all IDS operators cause the peak frequency of the MTF, and also any high-frequency cutoff (visual acuity), to increase proportionally with the square root of the mean luminance level. These increases continue up to luminance levels at which saturation begins to occur, i.e., the point-spread area shrinks to the size of a single receptor.

Psychophysical evidence indicates that the peak frequency and the high-frequency cutoff of the human spatial CSF show similar increases with mean retinal illuminance below the photopic range, though in general the changes are smaller than those expected from an IDS model. A plot of log visual acuity versus log retinal illuminance based on the data of Schlaer[7] is quite well fitted by a straight line with slope 3/8 (instead of 1/2) up to about 5 Td, after which acuity levels off rapidly. The spatial CSF's of Van Ness and Bouman[10] show a peak frequency that increases by 0.8 log unit (instead of 1) as mean illuminance increases from 0.09 to 9 Td. Raising mean illuminance beyond this point produces smaller changes in the CSF peak, and above 90 Td it appears that the entire CSF becomes independent of the mean luminance level.

Another difference between the behavior of IDS models and psychophysical data is that human CSF's generally show a decrease in sensitivity at the peak frequency as mean luminance decreases,[10,11] whereas the IDS model MTF maintains a constant gain at its peak frequency.

Discrepancies between IDS-model predictions and psychophysical data obtained at photopic luminance levels are to be expected in view of the model's automatic saturation property. It is interesting to note that the signal-detectability argument given in Section 1 implies that reliable detection of contrasts of the order of 0.1–1% covering an area the size of a single photoreceptor requires a quantum catch of the order of 10^6–10^8 times absolute threshold, or approximately 10–1000 Td. Over the range 10–1000 Td, then, the visual system loses its need for spatial summation, and so the disappearance of an IDS mechanism through saturation would not be disadvantageous. In this connection it is worth recalling that rod saturation occurs in the same range.[13]

Discrepancies below the photopic range call for a different sort of reconciliation. One approach is to weaken the IDS model's assumption that the point-spread area varies inversely with quantum catch. In Section 5 we develop a generalized IDS model in which that area varies as a power function of the input intensity. This allows the model to predict visual acuity and peak-frequency changes with mean luminance more in line with empirical results. A second approach is to take into account the time required for a point-spread effect to disperse across the retina. When plausible assumptions about this are combined with the actual temporal conditions prevailing during CSF measurements, preliminary analysis indicates that the IDS model yields a rise in peak-frequency sensitivity with increasing mean luminance comparable with that exhibited by human CSF's.

The exact shape of the MTF of an IDS model depends on the form of its point-spread function, and so it is an interesting coincidence that for the Gaussian case the MTF [Eq. (9)] turns out to be the same one produced by Marr and Hildreth's linear DEL^2-G model of early visual processing.[6] In that model the image is convolved with the Laplacian of the Gaussian function $-(1/\sigma^2 2\pi)\exp[-(1/2)(x^2 + y^2)/\sigma^2]$, i.e., with the sombrero-shaped point-spread function

$$(1/\sqrt{\pi})^2(1/\sigma^2)^2[1 - (x^2 + y^2)/2\sigma^2]\exp[-(1/2)(x^2 + y^2)/\sigma^2].$$

The Fourier transform of that point-spread function is

$$4\pi^2(u^2 + v^2)\exp[-2\pi^2\sigma^2(u^2 + v^2)],$$

and so its MTF for one-dimensional sinusoidal gratings is

$$4\pi^2 f^2 \exp(-2\pi^2\sigma^2 f^2).$$

It follows that the Gaussian IDS model cannot be distinguished from a single channel DEL^2-G model by experiments that simply determine the shape of the CSF at any fixed mean luminance level. [Such experiments generally involve small contrast values, in the range 10% or less, so that the approximation in Eq. (9) is valid. For high contrast values the nonlinearity of the IDS model would become an important factor and could allow an experimental discrimination between the models.] Marr and Hildreth[6] show that a DEL^2-G filter is essentially indistinguishable from a difference-of-Gaussians filter of the sort used by Wilson and Bergen,[14] and so the same is true of single-channel linear models based on that filter.

4. RESPONSES TO TWO-DIMENSIONAL PATTERNS

Ricco's Law and Weber's Law

Ricco's law states that the detectability of a spot of light depends only on the product of its area and intensity. Experimentally, in human vision, this holds for spots up to a certain critical size—a size that decreases as the background intensity increases.[2,3] We show here that the IDS model implies that Ricco's law holds for spots of all sizes on a background field of zero intensity—in the sense that the peak value of the output to such an input is the same for all spots of the same shape that have the same product of area times intensity. On nonzero backgrounds it causes Ricco's law to hold (in the same sense) for spots up to a critical area that decreases as the background intensity increases. (The experimental fact that Ricco's law holds for only a limited range of areas even on a nominally dark background does not necessarily contradict the model, since the activity in real visual systems does not fall to zero in darkness.) The IDS model also predicts the types of configurational effect reported by Sakitt,[15] who found that two separated spots lying within Ricco's area do not yield perfect summation but instead require more total quanta for detection than a single spot in the same area.

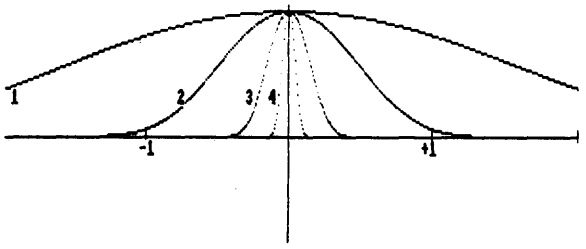

Fig. 8. Response profiles for square spots on a dark background. Spot area (A) times intensity (I) was held constant at 10. Curve 1, $I = 1$, $A = 10$; curve 2, $I = 10$, $A = 1$; curve 3, $I = 100$, $A = 0.1$; curve 4, $I = 1000$, $A = 0.01$.

Figure 8 shows the profiles of the Gaussian IDS-model response to square spots of various sizes on a zero background. The input image here was $I(x, y) = I$ for $|x| \leq W/2, |y| \leq W/2$, $I(x, y) = 0$ elsewhere (so the spot area was W^2). The output equation in this case is

$$\mathbf{O}(p, q) = \{N[(W/2 - p)\sqrt{I}] - N[-(W/2 + p)\sqrt{I}]\} \times \{N[(W/2 - q)\sqrt{I}] - N[-(W/2 + q)\sqrt{I}]\}. \quad (10)$$

In this figure all spots have a (area × intensity) value of 10. The response profiles shown here run along the horizontal axis through the center of the squares. It can be seen that the peak output value is the same for all inputs. This is a general property of IDS models.

Theorem 6

The peak value of the output to uniform patches of light on a zero-intensity background is the same for all patches of the same shape that have the same product (area × intensity).

Proof

For convenience we prove the theorem for square spots, but the form of the proof applies to any shape. Suppose that $I'(x,$

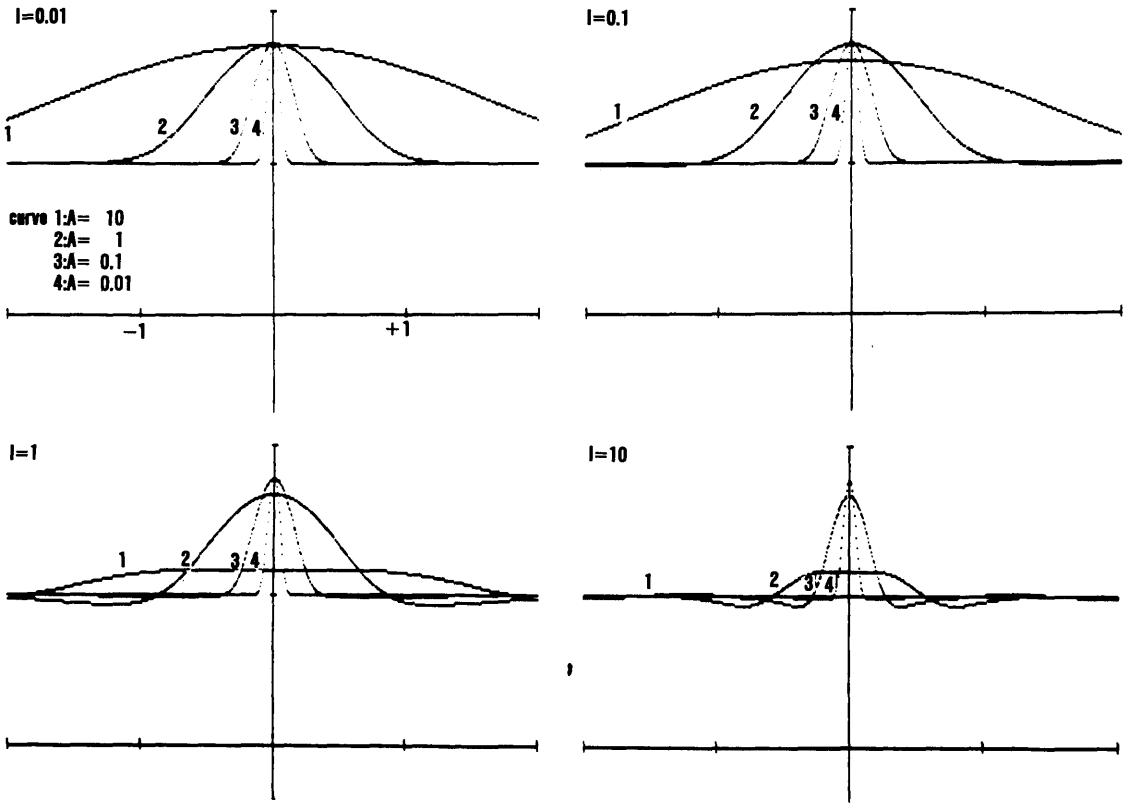

Fig. 9. Response profiles for square spots on nonzero backgrounds of various intensities. The input images were squares of intensity $I + D$ surrounded by backgrounds of intensity I. The spot area (A) times its incremental intensity (D) was held constant ($D \times A = 10$), and responses were computed for $A = 0.01$, 0.1, 1, and 10. Upper left, background intensity $I = 0.01$; upper right, $I = 0.1$; lower left, $I = 1$; lower right, $I = 10$.

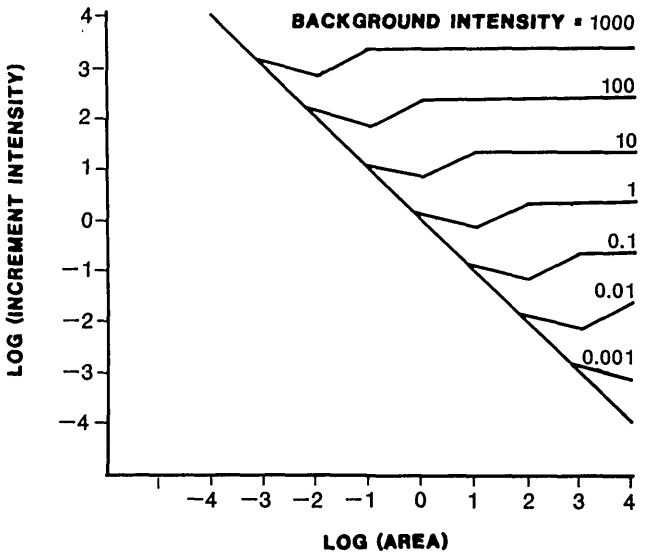

Fig. 10. Increment threshold as a function of test spot area for background fields of various intensities. The input images were square spots of area A and intensity $I + D$ surrounded by uniform background fields of intensity I. Each curve shows, in log–log coordinates, the value of D required to produce a peak response of 1.15 as A increases over eight log units. Background intensities range from $I = 1000$ (top curve) to $I = 0.001$ (bottom curve). The diagonal straight line represents Ricco's law; each curve follows this line up to some area value and then departs from it as shown.

y) is a square spot of width W and intensity I on a dark background, i.e., $I'(x, y) = I$ for $|x| \leq W/2, |y| \leq W/2$; and $I'(x, y) = 0$ elsewhere. And suppose that $I(x, y)$ is another square with intensity cI and width $W/\sqrt{c}$, so that (area × intensity) is $I \times W^2$ for both. Then $I(x, y) = cI'(x\sqrt{c}, y\sqrt{c})$, and so from Theorem 3

$$\mathbf{O}[I(x, y)](p, q) = \mathbf{O}(cI'(x\sqrt{c}, y\sqrt{c})](p, q) = \mathbf{O}[I'(x, y)](p\sqrt{c}, q\sqrt{c}).$$

Consequently, if the peak output to $I'(x, y)$ occurs at (p', q'), the peak output to $I(x, y)$ occurs at $(p'/\sqrt{c}, q'/\sqrt{c})$ and has the same value as the peak output to $I'(x, y)$.

For nonzero backgrounds, the IDS model implies that Ricco's law holds as an approximation for small spots: Up to a certain spot size the peak output value remains constant out to several decimal places (e.g., 3) for all spots (of the same shape) that have the same value of (area × intensity). The higher the background intensity, the smaller the critical area beyond which Ricco's law begins to fail.

Figure 9 shows the profiles of the Gaussian-model responses to square spots of various sizes on various backgrounds. Spot (area × intensity) was held constant at 10. On the lowest-intensity background (0.01) the peak-response value remains constant for areas ranging from 0.01 to 10. When the background intensity is increased to 0.1 the peak-response value is still constant for areas up to 1.0 but drops below the constant value for the largest spot (area = 10). For a background intensity of 1, only the two smallest spots preserve a constant peak output, and, finally, at the highest background intensity (10) Ricco's law fails for all but the smallest spot. (At this background intensity, Ricco's law would hold only for spots with areas ≤0.01.)

The equation for the Gaussian-model response to square spots of intensity $I + D$ on backgrounds of intensity I is

$$\begin{aligned}\mathbf{O}(p, q) = 1 + (&\{N[A(W/2 - p)] - N[-A(W/2 + p)]\} \\ &\times \{N[A(W/2 - q) - N[-A(W/2 + q]\}) \\ &- (\{N[B(W/2 - p)] - N[-B(W/2 + p)]\} \\ &\times \{N[B(W/2 - q)] - N[-B(W/2 + q)]\}), \qquad (11)\end{aligned}$$

where $A = (I + D)^{1/2}$, $B = \sqrt{I}$, and W is the spot width.

Figure 10 summarizes the Ricco law behavior of the model. It shows, for a range of background intensities, the spot intensity needed to produce a constant peak response as a function of spot area. (The spots here were squares, and the peak-response value at threshold was taken to be 1.15. That value was chosen for convenience: It is the peak response to a square of unit area when $D = 1$ and $I = 0.1$. The choice of threshold value is irrelevant here; other values yield curve families that look like those in Fig. 10.) For all background levels the constant-response curve runs for some distance along a line of slope −1, indicating obedience to Ricco's law, and then departs from this line when the spot area reaches a critical value. After a brief further decline with further increases in area (Piper's law), the curves increase a bit and then level out to constant values. For spot areas in that final range the peak response occurs as a Mach band at their edges and is governed by Weber's law.

Increment-Threshold versus Background-Intensity Curves

The last point is made more explicit by Figs. 11 and 12. Figure 11 replots three of the curves from Fig. 10 in the form of standard increment-threshold versus background-intensity (TVI) curves. It can be seen that these TVI curves evolve through three stages. When background intensity is low the curve is flat, as though threshold were limited by dark light (though here there is none). Next there is a transitional stage in which the TVI curve increases with a slope that is first somewhat less than one and then somewhat greater. Finally, when background intensity is sufficiently high, the TVI curve

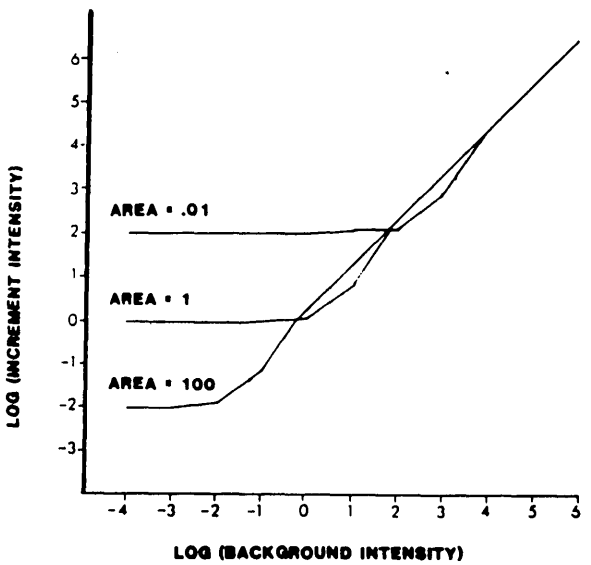

Fig. 11. TVI curves for test spots of different areas. These are replots of data from Fig. 10. Each curve shows the incremental intensity D required to produce a fixed peak-response value when the input is a square spot of area A and intensity $I + D$, surrounded by a background of intensity I. The three curves shown are for $A = 0.01$, $A = 1$, and $A = 100$. As background intensity increases, all curves eventually terminate in a diagonal straight line corresponding to Weber's law.

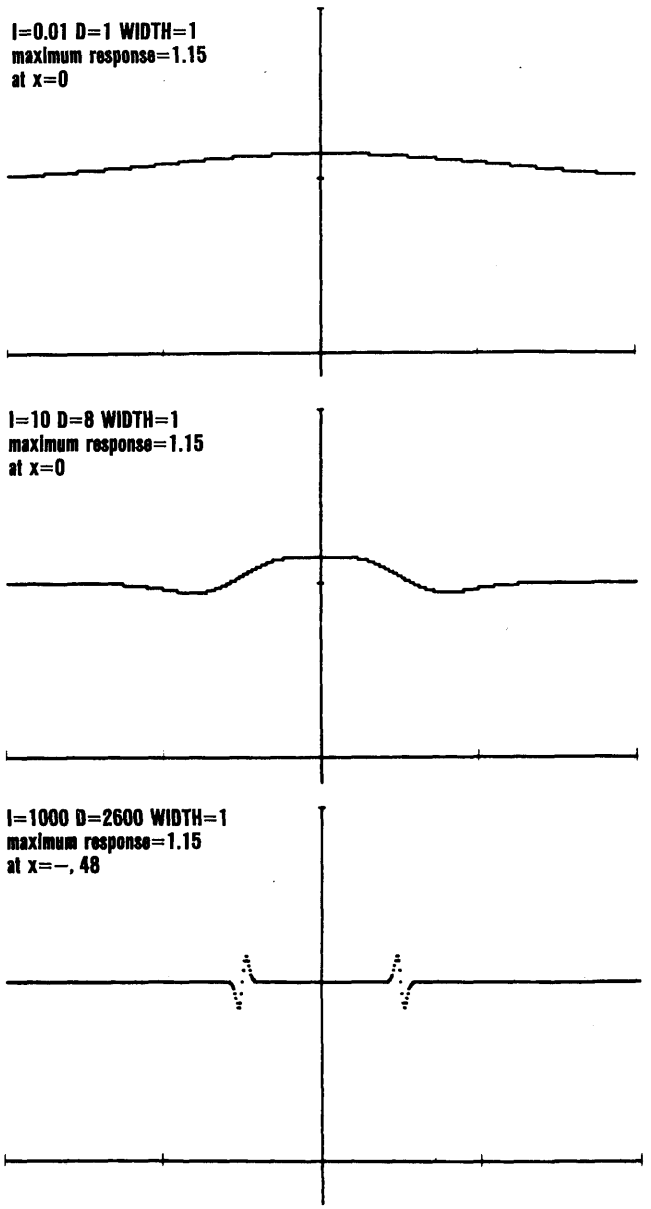

Fig. 12. Response profiles at threshold for a spot of fixed area on backgrounds of various intensities. Each curve shows the profile of the response to a square test spot of area $A = 1$ and intensity $I + D$ surrounded by a uniform background of intensity I. The increment value D in each case is that required to produce a peak output value of 1.15. Top profile, background intensity $I = 0.01$; middle, $I = 10$; bottom, $I = 1000$.

attains a slope of one (Weber's law) and retains it for all higher backgrounds. The background-intensity values corresponding to these three ranges depend on the size of the test spot: The larger the spot, the sooner its TVI curve begins to follow Weber's law.

These TVI curves are in good qualitative agreement with standard psychophysical results,[13] except that in the Weber's law region our curves all run together, whereas in practice one expects to find a slightly smaller threshold value of the Weber fraction for larger test spots.[16] This can be understood in terms of the fact that larger spots have longer perimeters, which should increase their relative detectability once the edge response becomes the dominant factor. We have not sought to model such an effect, since to do so realistically would introduce issues of noise and probability summation beyond the scope of this paper.

Figure 12 shows how the shape of the threshold-value response profile changes as background intensity increases. These profiles are for a test spot of area 1.0. On low-intensity backgrounds (in the zero-slope portion of the TVI curve) the response is simply a broad shallow bump, peaking in the center of the test spot. Here threshold is determined by the increment intensity required to make this central peak exceed the threshold criterion. In the next background-intensity range (corresponding to the transitional-slope portion of the TVI curve) the response profile at threshold has a sombrero shape, with apparent inhibitory regions surrounding a central positive bump. Here threshold is still determined by the response value at the center of the spot. Finally, on a high-intensity background, the response profile consists entirely of Mach bands at the edges of the test spot, and threshold is determined by their peak values. Those peaks follow Weber's law, as was shown earlier in Section 3, and this is the Weber region of the TVI curve.

Shape of the Impulse Response

Figure 13 illustrates, for small spots, a point made earlier for thin bars: At moderate to high background intensities, the IDS model produces a sombrero-shaped impulse response (center–surround antagonism), but when the same spot lies on a low-intensity background, the depression of surrounding activity becomes negligible, and the response appears to be

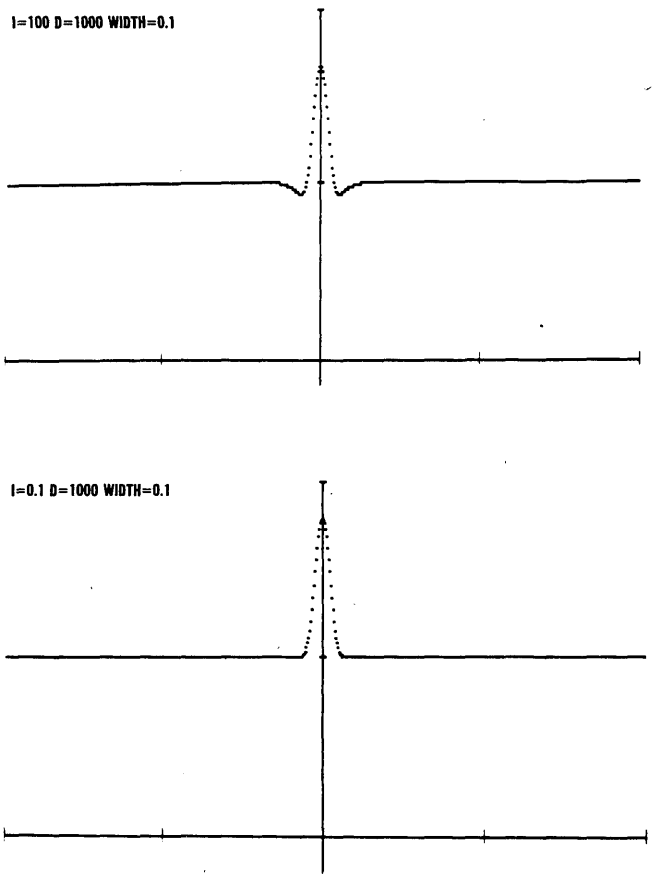

Fig. 13. Response profiles for a small square spot of fixed incremental intensity ($D = 1000$) superimposed upon a high-intensity (top curve, $I = 100$) or a low-intensity (bottom curve, $I = 0.1$) background. Spot width, 0.1.

purely positive—as though lateral inhibition had failed at low light levels. Mammalian retinal ganglion cells have been reported to behave in this fashion.[8,9] That fact has generally been interpreted in terms of a loss of the inhibitory contribution from the antagonistic-surround portion of the cell's receptive field. We see here that the same effect also occurs naturally in a system involving no inhibition.

The cause of this apparent loss of lateral inhibition on low-intensity backgrounds is that when the background level is low, the width of the point-spread function in the background region is large, and consequently the value of the response at points near the test spot is the sum of many small contributions coming from a large portion of the field. The high-intensity test spot reduces the amount of spread coming from receptors directly beneath it, but these are relatively few in number, and consequently their overall point-spread contribution to the response at nearby points is negligible to begin with. Thus, when it is removed, there is only a negligible reduction in the response level. When the background intensity is high, however, the point-spread function is narrow, and the response level at points near the test spot is the sum of spread values contributed by a relatively small number of closely neighboring points. In this case the loss of the spread values formerly contributed by points beneath the test spot causes a substantial reduction in the response level at points adjacent to that spot. Thus the same test spot creates appreciable "lateral inhibition" at nearby points when it is superimposed upon a high-intensity background and no apparent inhibition when the background is low.

Configurational Effects

At any given background intensity, Ricco's area can be defined as the area of the largest spot for which Ricco's law holds. If Ricco's law were the result of summation within the central region of a classical receptive field, one might expect all targets smaller than Ricco's area to be equally detectable if they have the same value of the product (area × intensity). Sakitt found, however, that Ricco's law is violated within Ricco's area when the target is a pair of spatially separated spots rather than a single continuous one.[15] Her experiment showed that two spots that deliver a fixed total number of quanta within Ricco's area may be undetectable even though the same number of quanta are detectable when imaged in the form of a single spot. Moreover, she showed that her results could not be reconciled with the idea of spatial summation over a fixed-size receptive field even if one allows for the possibility that receptors have different weights depending on their positions within the field.

For the IDS model these configurational effects pose no difficulty. It predicts what Sakitt found: The peak response to two spatially separated spots, each of area A and intensity/unit area D, is less than the peak response to one spot of area A and intensity $2D$, even though they lie entirely inside an area that would yield apparently perfect spatial summation when tested with larger continuous spots. Figure 14 illustrates this effect.

The top panel shows the response profile for a single square spot of intensity $I + D$ surrounded by a background of intensity I. I here is 0.1, and Fig. 9 shows that at this background intensity the width of Ricco's area is 1.0. The spot whose response profile is shown here has a width of 0.1, and its (area × intensity) value is 10. (That is, D is 1000.) The peak-response value for this spot is 1.78. The bottom panel shows the response profile for a pair of square spots, each of width 0.1, whose edges are separated by a gap of 0.05. The background intensity is again $I = 0.1$, and each spot has an intensity $D = 500$, so the combined (area × intensity) value for the two spots is 10. Thus this pair of spots falls well within the area of perfect spatial summation for this background intensity and have the same total (area × intensity) value as the single spot. However, the peak-response value for the pair is only 1.55.

This behavior can be understood qualitatively in the same way as the IDS model's creation of Mach bands at edges. Here the single spot's response contains a substantial contribution coming from receptors lying under the background portion of the input image. The responses to the separated squares gain a smaller contribution from spreading, because each square has a high intensity and consequently creates a narrower spread function in the receptors beneath it than they would produce if the low-intensity background were present. Thus each square reduces the point spread that its receptors would have contributed to the output of its neighbor.

The following expression is the output image equation for the Gaussian case of the IDS model when the input image is

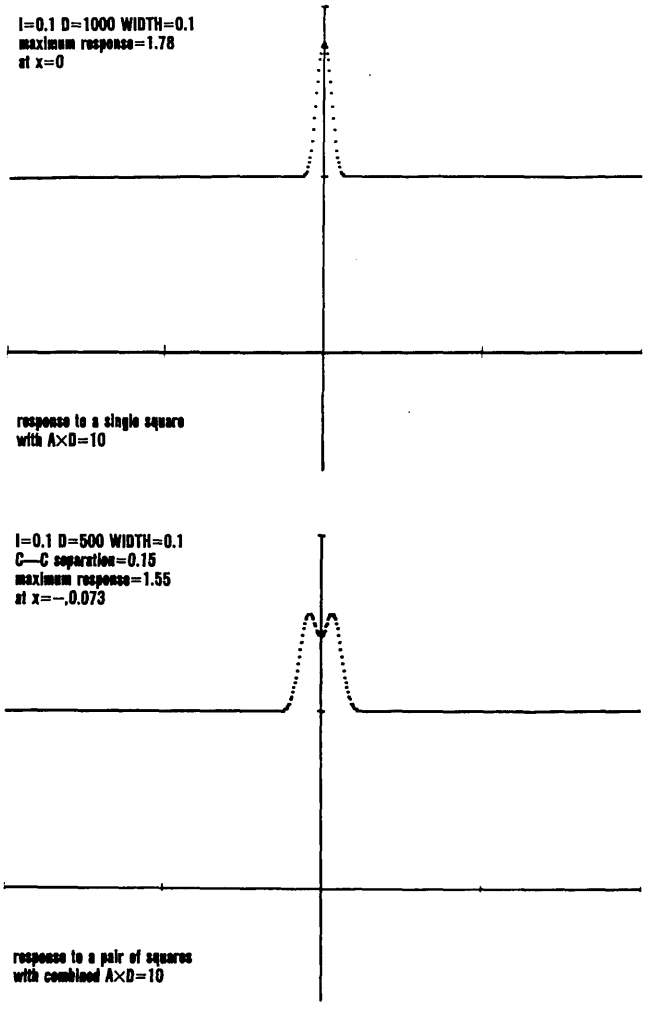

Fig. 14. Configurational effects within Ricco's area. The top curve is the response profile for a single square spot with (area × intensity) = 10. The bottom curve is the response profile for a pair of square spots whose combined (area × intensity) value was also 10.

a pair of squares of intensity $I + D$, width W, and center–center separation S, surrounded by a uniform background of intensity I. It assumes that the squares are both centered on the x axis. The curve in the bottom panel of Fig. 14 is a plot of the profile of this output function along the p axis (i.e., the horizontal axis of the output image).

$$\begin{aligned}O[p,q] = 1 &+ (\{N[A(W/2-q)] - N[A(-W/2-q)]\} \\ &\times \{N[A(S/2+W/2-p)] - N[A(S/2-W/2-p)]\}) \\ &+ (\{N[A(W/2-q)] - N[A(-W/2-q)]\} \\ &\times \{N[A(-S/2+W/2-p)] - N[A(-S/2-W/2-p)]\}) \\ &- (\{N[B(W/2-q)] - N[B(-W/2-q)]\} \\ &\times \{N[B(S/2+W/2-p)] - N[B(S/2-W/2-p)]\}) \\ &- (\{N[B(W/2-q)] - N[B(-W/2-q)]\} \\ &\times \{N[B(-S/2+W/2-p)] - N[B(-S/2-W/2-p)]\}),\end{aligned} \tag{12}$$

where $A = (I + D)^{1/2}$ and $B = \sqrt{I}$.

5. DISCUSSION

Intensity-Dependent Spatial Summation as a Psychophysical Model

For a model based on a single assumption, the IDS model gives a surprisingly complete first-approximation description of human spatial vision for retinal illuminances ranging from absolute threshold up to around 10 Td. It predicts the two major effects usually associated with spatial summation: the dependence of Ricco's area on background luminance and the fact that visual acuity increases approximately as the square root of mean luminance. And, unexpectedly, it also predicts two major effects that are not usually thought of as related to spatial summation—or, indeed, to each other: Mach bands and Weber's law. Those two effects are typically explained in terms of mechanisms quite different from the one embodied in the IDS model: lateral inhibition for Mach bands and nonlinear transduction for Weber's law. Here we examine the relationship between those familiar concepts and the IDS mechanism. We also describe a way in which the IDS model can be modified to produce a closer fit to psychophysical data and point out a connection between IDS processing and brightness constancy.

Mach Bands and Constant-Volume Models

Mach bands are generally attributed to a neural process of lateral inhibition that can be modeled by convolving the retinal image with a sombrero-shaped point-spread function whose negative brim represents the inhibition.[17] We will refer to this as the standard linear lateral inhibitory (LLI) model. Within the framework of linear systems theory, lateral inhibition is the only possible explanation of Mach bands, since Mach bands correspond to a high-pass filter effect and in a shift-invariant linear model such an effect can be produced only by a point-spread function containing negative lobes. However, we have seen that the IDS model, which is nonlinear, creates Mach bands with a purely positive point-spread function.

Thus the IDS model represents a new principle for generating edge enhancement, namely, edge enhancement will be produced by any model in which each photoreceptor creates a point-spread function whose volume is the same for all input intensities. Recall that the fundamental assumption of the IDS model is that the height of the receptor output function varies directly with input intensity but its volume remains constant. As a consequence, the effect of an image on the system is not to change its total output but rather to redistribute that output in space. It follows that, when the input is a uniform field, the output must also be uniform and that output level will be the same regardless of the input level—this is the intuitive proof that was given for Theorem 1 in Section 2. In other words, the sensitivity of the IDS model to uniform fields is zero.

Put another way, the IDS model has zero sensitivity at spatial frequency zero. And by extension it is clear that the same is true of any model in which the volume under the receptor output function remains constant across all input intensities. Furthermore, if the model responds at all, its sensitivity will rise from zero as frequency increases, so that it will act like a high-pass filter. And that, in turn, is what is generally meant by edge enhancement: Low frequencies are attenuated more than high frequencies, so that in the image itself large uniform areas are attentuated more than edges. It follows that all constant-volume models will produce edge enhancement.

An example of a constant-volume model different from the IDS model is illustrated in Fig. 15. Here the receptor point-spread function is the sum of two functions: a Gaussian whose variance remains constant and whose height is directly proportional to the input intensity I, added to another Gaussian whose variance also remains constant but whose height varies as $(c - I)$, so that the total volume under the spread function (i.e., the volume under the sum of the two Gaussians) is always equal to c regardless of the input intensity I. Because the volume is constant, this model will attenuate low frequencies and produce Mach bands. [Note that if c is positive the composite spread function will be entirely positive when the input level I is low and then will assume a sombrero shape at higher input levels, when $(c - I)$ becomes negative.]

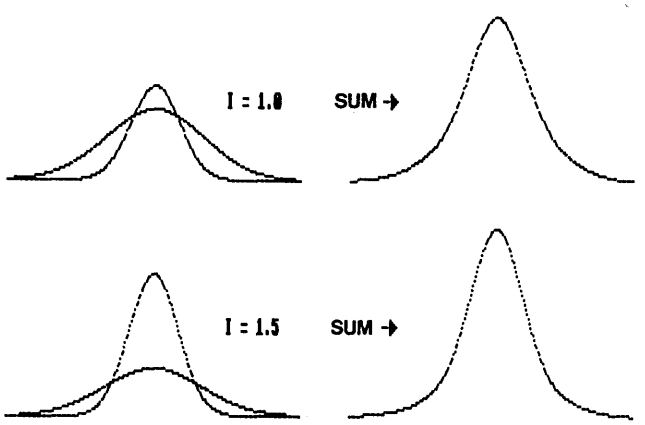

Fig. 15. Component curves for the point-spread function of a constant-volume model that differs from both the IDS model and the LLI model. Here the point spread is the sum of two functions, one whose height increases proportionally with the input intensity I [here, a Gaussian of the form $I \times G_1(x, y)$, where G_1 has a fixed standard deviation σ_1] and another whose height varies as $c - I$, where c is a positive constant [here, $(c - I) \times G_2(x, y)$, where G_2 has fixed standard deviation σ_2 and $\sigma_2 > \sigma_1$]. Component curves for two values of I are shown on the left, and the corresponding composite point-spread functions are shown on the right.

In general, this constant-volume model is nonlinear since superposition fails: The output to a uniform field with intensity $2I$ (i.e., c) is not twice the output to a field with intensity I (also c). However, when c equals zero the model is linear—in fact, it is the standard LLI model. Thus that model falls in the intersection between two distinct classes of model for edge enhancement: It is simultaneously a constant-volume model and a linear model with negative lobes in its impulse response. Since no linear model can be a constant-volume model unless the volume under its impulse response is zero (and consequently the value of its MTF is zero at the origin), it follows that the only linear models that can produce edge enhancement with complete dc suppression are also constant-volume models.

We see then that the edge-enhancement properties of the standard LLI model need not necessarily be attributed to inhibition *per se*. Instead, they could equally well be said to follow from the fact that it, like the IDS model, is a constant-volume model.

Weber's Law and a Generalized Intensity-Dependent Spatial-Summation Model

Weber's law is often attributed to an early nonlinear transformation in the visual system that causes the neural response to an input of intensity I to be approximately proportional to $\log I$. This very old idea is not really satisfactory because it does not explain why Weber's law fails at low luminances, and, more critically, even when buttressed with the concept of dark light, it still cannot explain why the range of background luminances for which Weber's law holds exactly should depend on the size of the test spot. The IDS model accounts for Weber's law and its failures on a totally different principle. Here the height of the receptor response varies linearly with the input intensity, and Weber's law arises as an edge effect that is due to spatial summation—an effect that begins to become significant only at a critical level of background luminance, which increases as the size of the test spot decreases. A natural question here is: What specific feature of the IDS model causes Weber's law to occur at all?

The answer cannot be the constant-volume assumption *per se*, since that assumption is shared by the standard LLI model, which does not imply Weber's law. And for the same reason, it cannot be the assumption that the height of the receptor spread function is directly proportional to the input intensity. In fact, the key to the model's Weber-law behavior is the constant-shape assumption, i.e., the assumption that the form of the spread function when the input intensity is I is $I \times S(Ir^2)$, where r is distance from the receptor. This assumption keeps the volume under the spread function constant by causing the equivalent area (volume/center height) to vary inversely with I. But closer analysis shows that this specific area-intensity relationship is not necessary for Weber's law. In fact, if the spread function takes the form $I^n \times S(I^n r^2)$, where n is any nonzero exponent, and all the other assumptions of the IDS model remain the same, the resulting model still implies Weber's law, though now the area under the spread function varies inversely with I^n instead of simply I. Thus the critical feature is really the fact that the model causes the spread function to change with intensity by rescaling the x and y coordinates of the retinal plane by factors that exactly undo the change in its height, thereby leaving its volume constant.

To prove that the generalized IDS model mentioned in the last paragraph implies Weber's law, recall that the key to our proof that all IDS models imply Weber's law, regardless of the form of the basic spread function (Theorem 4 in Section 3) was the scaling theorem (Theorem 3). Suppose now that we alter assumption (3) of Section 2 so that the spread to output point (p, q) from input point (x, y) with input intensity $I(x, y)$ is

$$(3A) \quad [I(x,y)]^n S\{[I(x,y)]^n[(x-p)^2+(y-q)^2]\},$$

where S is any spread function satisfying assumptions (1)–(4) of Section 2. And suppose that the output image is still the sum of the spread functions, i.e., assumption (5) now becomes

$$(5A) \quad \mathbf{O}[I(x,y)](p,q) = \iint [I(x,y)]^n \times S\{[I(x,y)]^n[(x-p)^2+(y-q)^2]\}dy\,dx.$$

Then Theorem 3 can be generalized as shown below.

Theorem 3A (Generalized Scaling Theorem)
For every positive constant c and every input image $I(x, y)$

$$\mathbf{O}[cI(x,y)](p,q) = \mathbf{O}[I(x/\sqrt{c}^n, y/\sqrt{c}^n)](p\sqrt{c}^n, q\sqrt{c}^n). \quad (13)$$

Proof
As in the proof of Theorem 3, we express the right-hand side of Eq. (13) in terms of the integral in assumption (5A) and make the change of variable $u = x/(\sqrt{c}^n), v = y/(\sqrt{c}^n)$. The result is the left-hand side of Eq. (13) expressed in integral form.

From Theorem 3A it is easy to prove that Theorem 4 still holds for this generalized IDS model, i.e., the maximum and the minimum values of the output on the high and the low sides of an edge still depend only on the ratio between the input intensities on the two sides. In other words, the generalized IDS model in which the point-spread area varies inversely with I^n still implies Weber's law in the same way as the original model.

In fact all the theorems proved for the original model still hold for this generalization, since their proofs in every case depended only on the scaling theorem. The only difference is that, wherever the original theorems and proofs mention the mean luminance level L, one needs to substitute L^n in the general case. Thus Theorem 5, which showed that visual acuity increases as $\sqrt{L}$, can be immediately generalized to show that acuity in this model increases as $\sqrt{L^n}$. As noted in Section 3, n values less than 1.0 are more in line with psychophysical acuity measurements (e.g., $n = 0.75$ for the data of Schlaer[7]). This is also true of measurements of the size of Ricco's area as a function of background luminance: Barlow's[2] results obtained at 6.5-deg eccentricity require a n of the order of 0.2, and the foveal data of Glezer[3] are fitted by $n = 0.5$.

Discounting the Illuminant: Weber's Law and Brightness Constancy

Most objects in natural scenes emit no light of their own but simply reflect light from the sun or some artificial source. Normally the reflectances of objects remain constant over time, but their illumination may vary by factors as large as 10^{10}, so the irradiance of their optical images can vary by the

same factor. After IDS processing the peak and trough amplitudes of the Mach bands at edges depend only on the ratio between the input image intensities on the two sides (Fig. 3). This ratio depends only on the reflectances of an object and its background and is independent of scene illumination. The shapes and the positions of these peaks and troughs, however, depend on the absolute input intensities and thus on illumination: Both become narrower and move closer to the edge itself as illumination increases. For any object–background combination, then, there is some illumination level beyond which the Mach bands generated on opposite sides of the object no longer overlap one another. At this level and all higher ones, the output image of the object consists of an edge-enhanced border whose peak and trough amplitudes depend on the reflectance ratio across its edges and whose interior has the baseline output value (1.0 for the IDS operators defined in Section 2). Of course this critical illumination level is lower the larger the object. Assuming that an object can be detected when the peak of its edge response differs from the baseline response value by more than some criterion amount, it follows that in an IDS system the detectability of any object will follow Weber's law once the illumination level gets high enough.

If the apparent brightness of an object is unaffected by its illumination and depends only on its reflectance and that of its background, as is roughly true in human vision, one speaks of brightness constancy. In the human visual system, the apparent brightness of the interiors of large objects of uniform luminance must be based on an extrapolation from their edges, since the retinal images of the interiors are effectively stabilized images and consequently cannot contribute to their visibility.[18] If an extrapolation mechanism based its assignment of interior brightnesses on the peak and trough values of the Mach bands at the edges of objects and received its input from an IDS operator, it too would exhibit brightness constancy for all objects beyond a certain size.

Intensity-Dependent Spatial Summation as an Image-Processing Algorithm

Intensity-dependent spatial summation seems potentially useful as a first-stage image-processing operation for applications involving the same type of boundary conditions faced by the retina—applications in which the inputs are Poisson noisy images whose mean intensity levels [(quanta/pixel)/frame] can vary substantially from scene to scene (e.g., because of changes in illumination) and also within a single image (e.g., because of shadows). These conditions occur naturally for television pictures of real scenes illuminated by the sun.

Automatic Gain Control

The illumination falling upon natural scenes can vary over the course of a day by as much as 10^{10}. No recording medium can readily accommodate such an enormous dynamic range. There are two fundamental objections to the usual solutions to this problem, such as the use of filters or amplifier gain changes. First, they are insensitive to local variations in scene illumination, e.g., owing to shadows: The effective luminance of the entire scene is reduced by a common factor, which can reduce the signal level in shadowed areas down into the range of the system noise. This is symptomatic of the second objection, which is more general. Spatial contrast detection is in principle limited by photon noise at all illumination levels; contrast sensitivity can always be improved by increasing the quantum catch. Thus any gain-control mechanism that simply enforces a fixed quantum catch, as the use of an iris or a filter does, is bound to become increasingly inefficient as the illumination level rises.

The IDS mechanism automatically compresses all input intensities into a output range extending from zero up to around twice the value of the constant point-spread volume (i.e., 0–2 when that volume is taken to be 1.0, as it was arbitrarily in the IDS model of Section 2.) In doing this it makes efficient use of every photon: As the image plane illuminance increases, the extra photons serve to decrease the size of the spatial-summation area, improving spatial resolution while maintaining a fixed reliability of contrast detection. And this effect occurs locally within a single image, so that in every region the size of the summation area is matched to the illumination falling upon objects in that portion of the scene.

Noise Smoothing and High-Frequency Attenuation

In noise smoothing by local averaging, the size of the summation area is usually held constant throughout any single image. The effect is simply low-pass linear filtering. This is a sensible way of suppressing photon noise, provided that the mean intensity level is known in advance (so that the summation area can be set inversely proportional to it) and that there is not much variation around the mean level within any single image. If the last condition cannot be guaranteed, either summation over a fixed area loses potentially resolvable high frequencies in the high-intensity regions of the image (because the summation area is too large for the mean luminance level in those regions), or else the low-intensity parts of the image become needlessly noisy (because the summation area is too small for the mean luminance level there), or both effects occur at once in different parts of the image.

The IDS operation, on the other hand, acts like a spatial filter whose high-frequency cutoff is always adjusted to match the prevailing light level (Fig. 7). In effect, it selects for attenuation the spatial frequencies that are so high, relative to the mean quantum catch/pixel, that they could not be reliably discriminated from photon noise. Thus the mean luminance level does not have to be known in advance, because the IDS mechanism adjusts to it automatically. And since this process occurs locally, different parts of the same image can have different mean intensity levels without requiring the mechanism to compromise on a single high-frequency cutoff. Instead, each region's cutoff frequency is automatically matched to its local mean intensity level. Thus, if the input is an image of a natural scene illuminated by the sun and some parts of the scene are in shadow, all parts of the output image will simultaneously tend to contain the maximum amount of high-frequency information justified by their local mean luminance levels.

Edge Enhancement

Edge enhancement is usually accomplished by convolving the input image with a more or less sombrero-shaped point-spread function consisting of a positive central region and a negative surround. For Poisson noisy optical images, this bandpass-filtering operation has no effect on the signal-to-noise ratio: If the input image takes the form $I \times r(x, y)$, where I is scene illumination and $r(x, y)$ is the reflectance distribution over a scene, after convolution the mean to standard deviation ratio

at each point is still proportional to $\sqrt{I}$. If the volume of the point-spread function is zero, as it usually is, uniform regions in the input image at any intensity I are converted into bandpass-filtered Gaussian noise with mean zero and variance I at every point. This noise is the background against which objects must be detected. For any value of I, the size of the sombrero must be adjusted to ensure an adequate signal-to-noise ratio at the Mach bands produced at edges, since those are the only places where most objects will be visible. In general, the critical size varies inversely with I, and, if the filter is poorly matched to the actual value of I in a given scene, the result will be either a needless loss of high-frequency information (when the sombrero is wider than necessary) or edges that cannot be discriminated from noise (when the sombrero is too small). If I varies greatly within a scene, the filter cannot be appropriate for all regions simultaneously, and one defect or the other is inevitable, just as with linear noise-smoothing filters.

An IDS operator acts like a bandpass filter whose frequency range automatically changes to match the prevailing mean-luminance level, both from scene to scene and also locally within scenes. Consequently, the parameter of an IDS filter (i.e., the width of its point-spread function) needs to be adjusted only for a single luminance level, and the filter will then adapt to all other levels (up to its saturation point), maintaining essentially the same size edge response at all levels for constant-contrast edges (because of the Weber-law property discussed above) and increasing spatial resolution as scene illumination increases. It can be shown that, for the Gaussian case, the IDS response to Poisson noisy uniform fields has a constant mean and variance for all values of $I \geq 0.01$. Consequently, the background noise against which objects are detected does not increase with scene illumination, and the detectability of edges (and thus of large targets) should remain constant as illumination increases, while resolution improves.

6. SUMMARY

We have analyzed a nonlinear model of retinal image processing, the IDS model, based on a single assumption: The height of the point-spread function varies directly with illuminance, whereas its volume remains constant, so that the area under the spread function around each photoreceptor is inversely proportional to the illuminance at that receptor. This assumption allows reliable spatial contrast discrimination in the face of photon noise while simultaneously maximizing spatial resolution. It proves to have the following consequences:

(1) *Bandpass Filtering.* The input image is effectively bandpass filtered, producing Mach bands at edges and an apparent center–surround antagonism in the response to small spots. In general, the model mimics effects normally attributed to lateral inhibition. This mimicry includes the fact that the appearance of lateral-inhibitory effects depends on illumination: At low background intensity levels, responses to small test spots exhibit no noticeable surround antagonism.

(2) *Ricco's Law.* For spatially continuous targets smaller than a critical size, the peak response value depends only on the product of target area times intensity. Thus detection of such targets should obey Ricco's law. The size of the critical area (that is, the size of Ricco's area) varies inversely with the background illuminance.

(3) *Configurational Violations of Ricco's Law.* Within Ricco's area (that is, the area of apparent perfect spatial summation as determined with spatially continuous targets), Ricco's law fails for noncontinuous targets: A single spot produces a larger peak response than two separated spots that have the same combined area × intensity product.

(4) *Del²-G MTF.* The response to low-contrast sinusoidal gratings closely approximates a sinusoid, allowing one to define a MTF. For the Gaussian case of the IDS model, the MTF at any fixed mean luminance level has the same form implied by a LLI model based on the negative Laplacian of a Gaussian.

(5) *Visual Acuity Improves with Illumination.* The MTF varies with illuminance in such a way that any high-frequency cutoff increases as the square root of the mean luminance level (for the simplest version of the model). This implies that visual acuity should vary in the same way.

(6) *Weber's Law Succeeds or Fails Depending on Target Size and Background Intensity.* The response to edges separating large uniform fields obeys Weber's law: The peak and trough values of the Mach bands at edges depend only on the ratio between the input image intensities on the two sides of the edge. When a target of fixed size is superimposed upon background fields of increasing intensity, its response profile evolves through three stages: first a simple bump, then a sombrero, and, finally, a pair of Mach bands at both edges with a baseline-response level between. The smaller the target is, the higher is the background level required to reach this final stage. Once it is reached, the detectability of the target satisfies Weber's law for all higher background luminance levels. In general, the model implies threshold versus background intensity curves whose shapes closely resemble those found in psychophysical experiments.

(7) *Brightness Constancy.* Assuming that the brightness of a target depends on the size of its edge response, the Weber property implies that sufficiently large targets will exhibit brightness constancy; i.e., their brightnesses will be independent of the scene illumination and depend instead only on their reflectances relative to that of the background.

APPENDIX A: DERIVATION OF THE RESPONSE TO LOW-CONTRAST SINUSOIDAL GRATINGS

We derive here the approximation given in Eq. (7). Suppose that the input is a vertical sinusoidal grating of the form $I(x) = 1 + k \cos 2\pi fx$. Then from Eq. (4) the output profile along the horizontal axis is exactly

$$O(p) = \int_{-\infty}^{\infty} [(1 + k \cos 2\pi fx)^{1/2}/\sqrt{(2\pi)}] \times \exp[(-1/2)(1 + k \cos 2\pi fx)(x - p)^2]dx.$$

For arbitrary values of k this integral seems quite intractable. However, when k is small enough that k^2 can be treated as zero, it can be solved as follows. First, write $1 + k \cos 2\pi fx$ as

$$[1 + (k/2)\cos 2\pi fx]^2 - (k^2/4)\cos^2 2\pi fx.$$

Dropping the second term, we have

$$(1 + k \cos 2\pi fx) \approx 1 + (k/2) \cos 2\pi fx,$$

and substituting this approximation into the output equation yields

$$\mathbf{O}(p) \approx \int_{-\infty}^{\infty} [(1 + j \cos 2\pi fx)/\sqrt{(2\pi)}] \times \exp[(-1/2)(x - p)^2] \times \exp[(-j)(\cos 2\pi fx)(x - p)^2]dx,$$

where $j = k/2$ and the factor $\exp[(-1/2)(k/2)^2(x - p)^2 \cos^2 2\pi fx]$ has been set equal to one. Expanding the second exponential factor as a Taylor series and dropping the terms containing powers of j greater than one, we have

$$\begin{aligned}\mathbf{O}(p) &\approx \int_{-\infty}^{\infty} [(1 + j \cos 2\pi fx)/\sqrt{(2\pi)}] \times \exp[(-1/2)(x - p)^2] \times [1 - j(x - p)^2 \cos 2\pi fx]dx \\ &= 1 - j \int_{-\infty}^{\infty} [1/\sqrt{(2\pi)}](x - p)^2 \cos 2\pi fx \times \exp[(-1/2)(x - p)^2]dx \\ &\quad + j \int_{-\infty}^{\infty} [1/\sqrt{(2\pi)}] \cos 2\pi fx \exp[(-1/2)(x - p)^2]dx \\ &\quad - j^2 \int_{-\infty}^{\infty} [1/\sqrt{(2\pi)}](x - p)^2 \cos 2\pi fx \times \exp[(-1/2)(x - p)^2]dx.\end{aligned}$$

Dropping the last term (which is less than j^2) and making the change of variable $v = x - p$, we obtain

$$\mathbf{O}(p) \approx 1 - j \int_{-\infty}^{\infty} [1/\sqrt{(2\pi)}]v^2 \cos 2\pi f(v + p) \times \exp[(-1/2)v^2]dv + j \int_{-\infty}^{\infty} [1/\sqrt{(2\pi)}] \times \cos 2\pi f(v + p) \exp[(-1/2)v^2]dv,$$

which can be solved exactly. Expanding the cosine factors into $(\cos 2\pi fv)(\cos 2\pi fp) - (\sin 2\pi fv)(\sin 2\pi fp)$ and noting that the integrals involving sine factors all vanish, we have

$$\begin{aligned}\mathbf{O}(p) &\approx 1 - j \cos 2\pi fp \int_{-\infty}^{\infty} [1/\sqrt{(2\pi)}]v^2 \times \cos 2\pi fv \exp[(-1/2)v^2]dv \\ &\quad + j \cos 2\pi fp \int_{-\infty}^{\infty} [1/\sqrt{(2\pi)}] \cos 2\pi fv \exp[(-1/2)v^2]dv.\end{aligned}$$

The third term can be obtained from integral tables: It works out to $j \cos 2\pi fp \exp(-2\pi^2 f^2)$. To evaluate the second term we note that the integral is the Fourier transform of $[1/\sqrt{(2\pi)}]v^2 \exp[(-\frac{1}{2})v^2]$, which is $[1 - (2\pi f)^2]\exp(-2\pi^2 f^2)$. The entire second term then is $-j \cos 2\pi fp$ times that expression. Combining all three terms and replacing j with $k/2$, we have finally

$$\mathbf{O}(p) \approx 1 + [2\pi^2 f^2 \exp(-2\pi^2 f^2)]k \cos 2\pi fp,$$

which is Eq. (7).

ACKNOWLEDGMENTS

Research for this paper was supported in part by NASA grant NCA2-OR345-301 to John I. Yellott, Jr. We thank A. Ahumada, S. Hersch, D. Kelly, and S. Reuman for discussions and assistance.

REFERENCES

1. Psychophysical demonstrations of spatial summation begin with A. Ricco, "Relazione fra il minimo angolo visuale e l'intensita luminoso," Ann. Ottalmol. **6**, 373–479 (1877); the literature is reviewed by P. E. Hallett, "Spatial summation," Vision Res. **3**, 9–24 (1963) and by B. Sakitt, "Configuration dependence of scotopic spatial summation," J. Physiol. (London) **216**, 513–529 (1971); Physiological demonstrations of spatial summation in the vertebrate retina begin with H. K. Hartline, "The receptive fields of optic nerve fibers," Am. J. Physiol. **130**, 690–699 (1940); a recent review of that literature is P. O. Bishop, "Processing of visual information within the retinostriate system," in *Volume III of the Handbook of Physiology: The Nervous System*, I. Darian-Smith, ed. (American Physiological Society, Bethesda, Md., 1984); spatial summation at the photoreceptor level (receptor coupling) was first reported by D. A. Baylor, M. G. F. Fourtes, and P. M. O'Bryan, "Receptive fields of cones in the retina of the turtle," J. Physiol. (London) **214**, 265–294 (1971); many studies are described in H. B. Barlow and P. Fatt, eds., *Vertebrate Photoreception* (Academic, New York, 1977).
2. H. B. Barlow, "Temporal and spatial summation in human vision at different background intensities," J. Physiol. (London) **141**, 337–350 (1958).
3. V. D. Glezer, "The receptive fields of the retina," Vision Res. **5**, 497–525 (1965).
4. When the illuminance is I, the quantum catch is a Poisson random variable with mean and variance IA, and when the illuminance rises to $I + cI$, the catch is Poisson with mean and variance $(1 + c)IA$. Taking the probability of detecting the change to be the probability that the catch for $I + cI$ exceeds the catch for I and using the normal approximation to the Poisson, it follows that the detection probability is the probability that a normal random variable with mean cIA and variance $IA(2 + c)$ is greater than zero. To make this probability greater than 0.999, $cIA/[IA(2 + c)]^{1/2}$ must be greater than 3. The order-of-magnitude value $10/c^2$ underestimates the actual required value of IA [i.e., $(9/c^2)(2 + c)$] by a factor ranging from 0.55 (when $c = 0.01$) to 0.37 (when $c = 1$).
5. A. Rose, "The sensitivity performance of the human eye on an absolute scale," J. Opt. Soc. Am. **49**, 645–663 (1948); *Vision: Human and Electronic* (Plenum, New York, 1974).
6. D. Marr and E. Hildreth, "Theory of edge detection," Proc. R. Soc. London Ser. B **207**, 187–217 (1980).
7. S. Schlaer, "The relation between visual acuity and illumination," J. Gen. Physiol. **21**, 165–188 (1937). [Schlaer's data are shown in Fig. 3 of J. P. Thomas, "Spatial resolution and spatial interaction," in *Handbook of Perception*, E. C. Carterette and M. P. Friedman, eds. (Academic Press, New York, 1975), Vol. V, Chap. 7.]
8. H. B. Barlow, R. Fitzhugh, and S. W. Kuffler, "Change of organization in the receptive fields of the cat's retina during dark adaption," J. Physiol. (London) **137**, 338–354 (1957).
9. C. Enroth-Cugell and J. G. Robson, "The contrast sensitivity of the ganglion cells of the cat," J. Physiol. (London) **187**, 517–552 (1966); A. M. Derrington and P. Lennie, "The influence of temporal frequency and adaptation level on receptive field organization of retinal ganglion cells in cat," J. Physiol. (London) **333**, 343–366 (1982). These experiments measured spatial CSF's for individual X cells over a wide range of mean luminance levels and fit them with modulation transfer functions implied by a linear difference-of-Gaussians receptive field model. In both cases the X-cell CSF changed from bandpass to low pass as mean luminance fell from photopic levels to near absolute threshold, indicating a loss of lateral inhibitory effects. The parameters of the best fitting MTF's implied that this change was due almost entirely to changes in the relative sensitivities of the center and surround mechanisms: The spatial areas of the center and surround apparently changed very little with mean luminance. Analyzing these data from a signal-detection standpoint, we find that that interpretation implies a very large decrease in the

quantum efficiency of the cat retina with light adaptation: for Derrington and Lennie's *X*-cell 25-J (their Fig. 9) quantum efficiency apparently fell by around 4 log units as mean luminance increased from 3.8×10^{-5} to 200 cd/m^2. Psychophysical evidence indicates that human quantum efficiency falls by only a factor of 10 over the same range (Ref. 5). Comparative visual-acuity measurements show that as mean luminance rises from 10^{-5} to 10 cd/m^2, human visual acuity improves by a factor of 30, whereas cat acuity rises by only a factor of 3. [T. Pasternak and W. H. Merigan, "The luminance dependence of spatial vision in the cat," Vision Res. **21**, 1333–1339 (1981)]. Taken altogether, these results suggest that cat and human retinas respond quite differently to changes in the light level. We are not aware of any study measuring spatial CSF's for primate retinal ganglion cells as a function of mean luminance, but we would expect substantial changes in the apparent size of receptive fields.

10. F. L. Van Ness and M. A. Bouman, "Spatial modulation transfer in the human eye," J. Opt. Soc. Am. **57**, 401–406 (1967).
11. D. H. Kelly, "Adaptation effects on spatio-temporal sine-wave thresholds," Vision Res. **12**, 89–101 (1972).
12. H. B. Barlow, "Increment thresholds at low intensities considered as signal/noise discriminations," J. Physiol. (London) 136, 469–488 (1957).
13. M. Aguilar and W. S. Stiles, "Saturation of the rod mechanism of the retina at high levels of stimulation," Opt. Acta **1**, 59–65 (1954).
14. H. R. Wilson and J. R. Bergen, "A four mechanism model for spatial vision," Vision Res. **19**, 19–32 (1979).
15. B. Sakitt, "Configurational dependence of scotopic spatial summation," J. Physiol. (London) **216**, 513–529 (1971).
16. B. H. Crawford, "Visual adaptation in relation to brief conditioning stimuli," Proc. R. Soc. London Sec. B **134**, 283–302 (1947). [Data shown as Fig. 3.10 in H. Ripps and R. A. Weale, "Visual adaptation," in *The Eye*, 2nd ed., H. Davson, ed. (Academic, New York, 1976), Vol. 2A.]
17. F. Ratliff, *Mach Bands: Quantitative Studies on Neural Networks in the Retina* (Holden-Day, San Francisco, Calif., 1965).
18. J. Krauskopf, "The effect of retinal image stabilization on the appearance of heterochromatic targets," J. Opt. Soc. Am. **53**, 741–744 (1963).