Model of visual adaptation and contrast detection

GEORGE SPERLING[1]
BELL TELEPHONE LABORATORIES

A three-component model of spatial vision is proposed, consisting of (1) a feedback stage, (2) a feedforward stage, (3) a threshold detector. The components correspond to physiological processes; in particular, the feedforward control signal corresponds to the "surround's" signal in the receptive fields of retinal ganglion cells. The model makes appropriate qualitative predictions of: (1) a square-root law ($\Delta\ell \propto \ell^{1/2}$) for detection at low luminances, (2) a Weber law ($\Delta\ell \propto \ell$) at high luminances, (3) additivity of threshold masking effects at high background luminances, (4) receptive fields that, in the dark, consist only of an excitatory center and that, in the light, also contain inhibitory surrounds, (5) the variation of spatial characteristics of receptive fields depending on the temporal characteristic of the test stimulus used to measure them, (6) the subjective appearance of Mach bands, (7) sine-wave contrast-threshold transfer functions, (8) the frequent failure of disk-detection experiments to demonstrate inhibitory surrounds, and (9) various second-order threshold effects, such as reduced spatial integration for long-duration stimuli, reduced temporal integration for large-area stimuli, and the increased effect of background luminance on the detection of large-area stimuli. Predictions are improved by assuming there exist various sizes of receptive fields that determine thresholds jointly.

I. INTRODUCTION

This article describes a shunting feedback-plus-feedforward model of visual contrast detection.[2] Shunting networks derive from neurophysiology, where they describe a kind of synaptic inhibitory process. Though nonlinear, the mathematical analysis of the steady-state response of the shunting networks is basically simple. On the other hand, human contrast detection is basically complex, probably involving several different mechanisms. Whereas a complete model for contrast would be correspondingly complex, the model proposed here is neither complex nor exact. The justification for proposing it is that, for a model of its simplicity, it has remarkable correspondences to vision.

Shunting Inhibition

Fatt and Katz (1953) and Coombs, Eccles, and Fatt (1955) first established the basic mechanism of neural inhibition. This mechanism may be called shunting inhibition (Furman, 1965) because inhibitory signals cause a portion of the excitatory signals to be diverted, or shunted. An electronic analog of shunting inhibition is an RC-stage in which the value of R is decreased by inhibition (Fig. 1). At high levels of inhibitory input, the net effect is analagous to arithmetic division of excitation by inhibition.

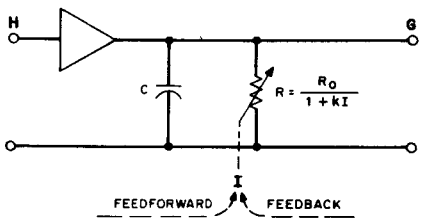

Fig. 1. Electronic analogue of shunting synaptic inhibition. The input current is H; the output voltage is G; the triangle indicates an isolating transconductance. The shunt path is R, which is controlled by inhibition, I. Inhibition may arrive either via feedforward or feedback paths; I increases the conductance (diminishes R) of the shunt path proportionally to its own strength.

Lateral Interaction

All visual systems of more than one receptor exhibit phenomena of lateral interaction; namely, the outputs of the adjacent receptors combine at various levels in the nervous system. A priori, at a particular level, lateral interaction may be characterized as feedback or feedforward, and as shunting or subtractive (Furman, 1965), although these categories are not exhaustive. For example, the lateral interaction in the eye of *Limulus* has been characterized as feedback and subtractive (Hartline & Ratliff, 1958).[3]

Various subtractive models of lateral interaction have been proposed for vision (e.g., Schade, 1956; von Békésy, 1960; Lowry & DePalma, 1961; Rodieck, 1965; Nachmias, 1968). A desirable feature of these models is that, because they are linear, the full power of linear analysis applies. Unfortunately, linear models do not handle the problem of light adaptation, except by ad hoc mechanisms. On the other hand, shunting models of lateral interaction intrinsically are models of adaptation.

Boundary Detection

When viewing the boundary between two adjacent, uniform fields of slightly different luminances, Os see an illusory light band near the boundary on the light side and an illusory dark band near the boundary on the darker side (Fig. 2). These illusory light-dark bands are called Mach bands because of their similarity to the light-dark bands observed by Mach at discontinuities in luminance gradients (Ratliff, 1965).

Because of Mach bands, the apparent contrast between two fields of nearly equal luminance is greatest at the boundary between them. Thus, in searching for a liminal test field superimposed on a masking field, it usually is the boundary of the test that the O detects (Lamar et al, 1947). For a boundary to be detected, it is assumed that the excitation in the light band (plus-zone, z_+) must differ by a criterion amount from the excitation in the

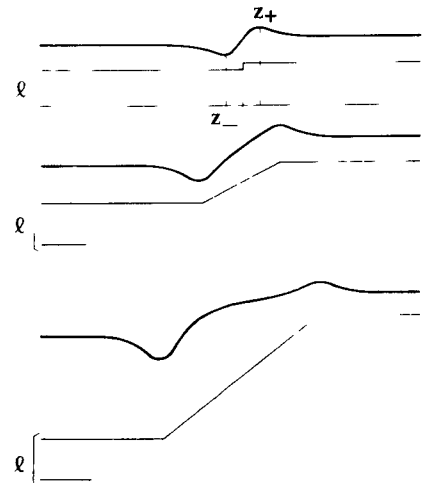

Fig. 2. Mach bands at a boundary (a), and at the ends of gradients (b) and (c). The angular curves indicate the stimulus retinal illuminance $\ell(x)$ as a function of x; the short horizontal lines indicate $\ell(x) = 0$. The smooth curves are G(x), the predictions by the model of the retinal response to $\ell(x)$. G(x) corresponds approximately to the subjective appearance of the stimulus. In (a), the light band z_+ and the dark band z_- are indicated.

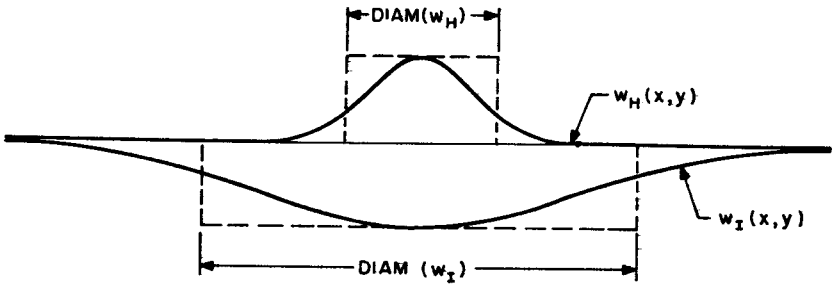

Fig. 3. Example of an excitatory weighting function $w_H(r)$ and of an inhibitory weighting function $w_I(r)$. Both functions are two-dimensional normal distributions; $\sigma_I/\sigma_H = 3$. These functions are used in all numerical examples of this article. The indicated effective diameters of w_H and w_I are $2\sigma_H\sqrt{2}$ and $2\sigma_I\sqrt{2}$, respectively.

dark band (minus-zone, z_-). An expression for the difference in excitation between z_+ and z_- is derived from the feedforward shunting model.

II. FEEDFORWARD MODEL

Outline

Let $\ell(x,y)$, the illuminance distribution on the retina, be the input to the model. The input undergoes a transformation (feedforward neural field) so that corresponding to each point x,y there ultimately is an output G(x,y). The output G is composed of an excitation term H and an inhibition term I. The excitation term H is given by a weighted sum of illuminances falling on the *excitatory area* around x,y; inhibition is given by the weighted sum of illuminance falling on a larger, concentric *inhibitory area* (Fig. 3). Inhibition interacts with excitation by feedforward shunting to produce the net output G. A detector scans G(x,y) to find the point z_+ where G(z) is a maximum and the point z_- where G(z) is a minimum. When ΔG,

$$\Delta G = \max G(z) - \min G(z)$$
$$= G(z_+) - G(z_-),$$

exceeds a threshold criterion ϵ, detection is signaled.

Adaptation

As a complication, the input to the feedforward neural field may not be $\ell(x,y)$ directly, but some transformation of ℓ. There are good reasons for supposing that the H and I terms of the feedforward field are not composed simply of weighted sums of ℓ, but that a feedback transformation (adaptation) precedes the feedforward transformation (Sperling & Sondhi, 1968). If the initial adaptation is taken to be the shunting feedback transformation proposed by Fuortes and Hodgkin (1964) and by Sperling and Sondhi (1968), then results obtained with pure feedforward shunting carry over—with only slight modification—to the case of feedback plus feedforward.

The next section presents a formal statement of the assumptions. Readers who are not interested in these details may skip to Section III.

IIa. Basic Assumptions

(1) *Excitation.* At each point in space x,y and at each instant in time t, excitation H is given by the weighted sum of retinal illuminance ℓ in a spatial and temporal region around x,y,t, namely

$$H(x,y,t) = \int_0^\infty \iint_R \ell(x',y',t')\, w_H (x-x',y-y')\, v_H(t-t')\, dx'\, dy'\, dt'. \quad (1)$$

Here w_H represents the excitatory area-weighting function, v_H represents the excitatory temporal-weighting function, and R represents integration over the area of the retina.

By defining the symbol * to denote the operation of convolution, Eq. 1 can be rewritten as

$$H(x,y,t) = \ell(x,y,t) * w_H(x,y) * v_H(t). \quad (1a)$$

The definition of H by an integral is an approximation to a sum that involves all the receptor elements; because these are numerous, the approximation is good.

(2) *Inhibition.* Inhibition I(x,y,t) is given by

$$I(x,y,t) = \ell(x,y,t) * w_I(x,y) * v_I(t) \quad (2)$$

where w_I and v_I are the inhibitory area-weighting and temporal-weighting functions. Figure 3 illustrates the w_H and w_I that are used in the examples of this article.

(3) *Output.* The output G(x,y,t) is defined by

$$\frac{d}{dt} G(x,y,t) + \frac{1 + kI(x,y,t)}{RC} G(x,y,t) = \frac{1}{C} H(x,y,t). \quad (3)$$

Equation 3 arises from the circuit of Fig. 1. which describes shunting inhibition. For H and I constant in time, Eq. 3 reduces to $G = R \cdot H/(1 + kI)$. For convenience, the parameter R is absorbed into G so that—for steady inputs—Eq. 3 reduces to the form

$$G = \frac{H}{1 + kI}. \quad (3a)$$

It is interesting to note that a similar equation for lateral interaction originally was proposed by Mach one hundred years ago (Mach, 1868; Ratliff, 1965).[4]

(4) *Prior adaptation.* The adaptation mechanism proposed by Fuortes and Hodgkin (1964) consists of a cascade of n RC stages in each of which the R is controlled by shunting feedback from the final stage. This n-stage feedback system is described by equations of the form

$$\frac{d}{dt} Y_j(\lambda) + Y_j(\lambda)[1 + Y_n(\lambda)] = Y_{j-1}(\lambda) \qquad j = 1, \cdots, n$$

where $\lambda = t/RC$ and $Y_0 = \ell$ (Sperling & Sondhi, 1968). For constant input ℓ, the output Y is defined by

$$\ell = Y(1 + Y)^n. \quad (4)$$

In the case of prior adaptation, the ℓ of Eqs. 1 and 2 is replaced by Y of Eq. 4.

(5) *Boundary detection.* The detector locates the point z_+ with maximum output in the light band near the boundary and the point z_- with minimum output in the dark band of the boundary. It forms the difference $\Delta G = G(z_+) - G(z_-)$, adds a random variable N to this difference, and produces a detection response if $\Delta G + N \geq \epsilon$. The criterion ϵ may vary from experiment to experiment but is assumed to remain fixed within an experiment.

IIb. Simplifying Assumptions; Corollaries

The aim of this article is to illustrate general properties of the model that obtain over wide ranges of parameters. We, therefore, make the following simplifying assumptions.

(1) *Homotropic.* The weighting functions $w_H(x,y)$, $w_I(x,y)$ are assumed

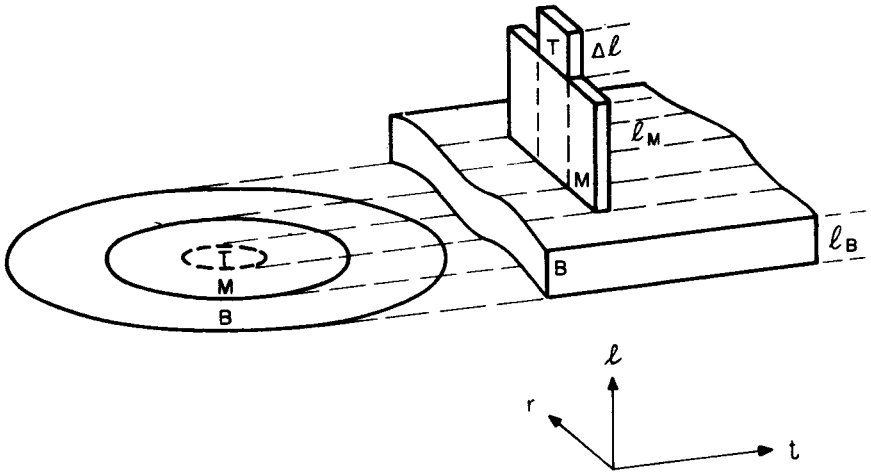

Fig. 4. Typical spatial and temporal relations between background (B), masking (M), and test (T) fields. Spatial relations only are illustrated at left; spatial and temporal relations at right. On the right, the ℓ-axis represents retinal illuminance ℓ, the t-axis represents time and the r-axis represents the diameter of the stimuli (after Sperling, 1965).

not to vary within the domain being studied. Further, they are assumed to be radially symmetric, that is, functions only of $r = \sqrt{x^2 + y^2}$. The weighting functions w_H, w_I are defined so that (a) the maximum value of each is 1, and (b) w_I extends beyond w_H. Assumption a is arbitrary for convenience; Assumption b is amply supported at the retinal level by physiological evidence.

(2) *Additivity of luminance components.* When the net retinal illuminance $\ell(x,y,t)$ is composed of several component illuminances, such as a background illuminance $\ell_B(x,y,t)$, a masking illuminance $\ell_M(x,y,t)$, and a test $\ell_T(x,y,t)$, then the contributions of each of these illuminance components to H and to I can be computed separately and the results added together. For example,

$$H = [\Sigma_j \ell_j(x,y,t)] * w_H(x,y) * v_H(t)$$

$$= \Sigma_j \ell_j(x,y,t) * w_H(x,y) * v_H(t) \quad (5)$$

where ℓ_j denotes a component illuminance. Equation 5 follows from the superposition (linear) property of convolution.

(3) *Effective areas and durations.* In order to determine whether or not a boundary is detected, it is not necessary to find ΔG as a function of time; it is only necessary to find its maximum value. Suppose ΔG attains its maximum at time t_m. By assuming the interaction to be of the form of Eq. 3a, the problem of calculating ΔG reduces to calculating H and I at t_m.

The calculations of H and I usually can be simplified by considering ℓ to be composed of several fields; a field to be detected—the test, T, a background field, B, and/or a masking field, M. B and M impair the detectability of T over the time and areas of interest. Usually M denotes a time-varying field $\ell_M(x,y,t)$. Typical relations between T, M, and B are illustrated in Fig. 4.

The contribution of a field j of retinal illuminance $\ell_j(x,y,t)$ to H at t_m is given by $H_j = \ell_j * w_H * v_H$. The units of H_j are total illuminance energy, i.e., retinal illuminance × area × time; it is useful to think of H_j as the total energy contributed by Field J to H at the moment of detection. To calculate H_j it usually is not necessary to know $w_H(x,y)$ and $v_H(t)$ exactly. For example, when $\ell_B(x,y,t) = \ell_B$, i.e., a constant, it only is necessary to know

$$W_H = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_H(x,y)\,dx\,dy,$$

the effective area of w_H, and

$$V_H = \int_0^{\infty} v_H(t)\,dt,$$

the effective duration of v_H (cf. Fig. 3). The energy contributed by B to H then is $H_B = \ell_B \cdot W_H \cdot V_H$.

When a field does not cover w or v entirely, estimates can be made of the intersection between the field and the weighting function, i.e., of their fractional overlap (Fig. 5). In the case of area weighting functions, good estimates can be made for the cases of interest (see below and Fig. 5). For the case of temporal weighting functions, two cases are by far the most frequent: (a) $\ell_j(x,y,t) = \ell_j(x,y)\,\delta(t)$ and (b) $\ell_j(x,y,t) = \ell(x,y)$. In Case 1, the temporal weight is 1.0; and in Case 2, only the effective areas (V_H, V_I) matter. By confining our treatment to T fields, which either are impulses or steady, to M fields, which are impulses coincident with T or steady, and to B fields, which are steady, only one parameter, V_I/V_H, need be known to describe the temporal weighting functions.

(4) *Comparison-detection of neighboring areas.* It may seem artificial that detection is based on the difference in output between two points, z_+, z_-, rather than between two areas, e.g., appropriately weighted neighborhoods of z_+ and z_-. Comparison of neighborhoods rather than points is equivalent to spatial filtering in the detector, that is, to a convolution integral of G(x,y) with the neighborhood weighting function of the detector. Since, in the model, the effects of spatial filtering at different levels are difficult to discriminate, all spatial filtering is assumed to occur in the feedforward neural field.

IIc. Derivation of Predicted Contrast Thresholds for Incremental Tests upon Uniform Backgrounds

Variation of incremental threshold luminance $\Delta\ell$ with background luminance ℓ. Approximate solution. The critical difference ΔG for detection is assumed to be

$$\Delta G = G(z_+) - G(z_-)$$

$$= \frac{H(z_+)}{1 + kI(z_+)} - \frac{H(z_-)}{1 + kI(z_-)}. \quad (6)$$

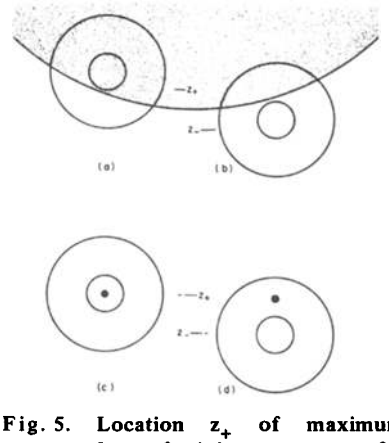

Fig. 5. Location z_+ of maximum response and z_- of minimum response for two stimuli: (a) and (b) a light-dark boundary; (c) and (d) a small test spot. The effective areas of w_H and of w_I are drawn around z_+ and z_-. Receptive fields are illustrated for (a) z_+ inside light areas; (b) z_- in dark area; (c) z_+ centered around stimulating spot; (d) z_- located with test spot at fringe of w_H but within w_I.

Defining $\Delta H = H(z_+) - H(z_-)$, $\Delta I = I(z_+) - I(z_-)$, substituting into Eq. 6, and algebraically reducing gives

$$\Delta G = \frac{\Delta H - G(z_-)k\Delta I}{1 + kI(z_+)} \cdot \quad (7)$$

Rearranging terms of Eq. 7 gives

$$\Delta H = \Delta G + k\Delta G\, I(z_+) + kG(z_-)\, \Delta I. \quad (8)$$

Equation 8 can be rewritten as

$$\Delta H = \Delta G + [k\Delta G]\, I(z_-) + [kG(z_+)]\, \Delta I. \quad (9)$$

To simplify Eq. 9 further, the following assumptions are made:

(1) M and B entirely cover the area of interest so that only T contributes to differences between z_+ and z_- in the amount of H and I received.

(2) ΔH can be expressed in terms of fractional overlaps as follows. Let $\Delta\ell = \max \ell_T(x,y)$. From the definition of H(x,y) by Eq. 1, it follows that $H_T(z_+)$ can be expressed as some constant $[p(z_+,T) \cdot W_H]$ times $\Delta\ell$, where $p(z_+,T)$ represents the fractional overlap of the area W_H with T at the point z_+. Similarly, $H_T(z_-) = p(z_-,T) \cdot W_H \cdot \Delta\ell$. To calculate ΔH, the additivity property (Eq. 5) is convenient:

$$\Delta H = H(z_+) - H(z_-)$$

$$\Delta H = [H_T(z_+) + H_{M+B}(z_+)] - [H_T(z_-) + H_{M+B}(z_-)].$$

Since it was assumed in (1) that M and B uniformly covered the entire area of interest, $H_{M+B}(z_+) = H_{M+B}(z_-)$ giving

$$\Delta H = H_T(z_+) - H_T(z_-)$$

$$= [p(z_+,T) - p(z_-,T)]\, W_H \cdot \Delta\ell$$

$$= \Delta p \cdot W_H \cdot \Delta\ell$$

where the constant Δp is defined as $p(z_+,T) - p(z_-,T)$. Thus, ΔH, the left side of Eq. 9, reduces to a constant $(\Delta p \cdot W_H)$ times $\Delta\ell$.

(3) The time-relations and geometry of M, T, and B are assumed to remain fixed and either (a) ℓ_M and ℓ_B are varied together ($\ell = c\ell_M + c'\ell_B$) or (b) only one of ℓ_M and ℓ_B (denoted ℓ is varied). By this definition of ℓ, $I(z_-)$ becomes $c'' + c'''\ell$, where the c's depend on the time relations and geometry of M, T, and B.

(4) For the moment, ΔI will be assumed to be negligibly small so that the last term of Eq. 9 vanishes. At threshold, ΔG is constant so Eq. 9 reduces to

$$\Delta\ell = c_1 + c_2\ell \quad (10)$$

where the positive constants c_1 and c_2 depend on the overlap of the B, M, and T fields with the weighting functions, on the threshold criterion ΔG, and on the relative amounts of ℓ_M and ℓ_B. This elementary equation describes the response of the feedforward system to the usual contrast detection stimulus, i.e., an incremental test field T of threshold luminance $\ell_T = \Delta\ell$ added onto a larger background field B of luminance ℓ. Variations in procedure cause predictable changes in the constants c_1 and c_2.

Variation of $\Delta\ell$ with ℓ. Complete solution. The exact solution of Eq. 6 is much more complicated than the approximate solution (Eq. 10), but it is basically quite similar to Eq. 10. To derive this solution, we may start with Eq. 8, expand $I(z_+)$ into $I(z_-) + \Delta I$, $G(z_-)$ into $H(z_-)$ and $I(z_-)$, and make the following substitutions, which will be explained later. For $H(z_-)$ substitute $I(z_-)/\sigma^2$; for ΔI substitute $a\sigma^2\Delta H$; for $k\sigma^2\Delta G$ write ϵ. This ultimately yields

$$\Delta H = \frac{\Delta G[1 + kI(z_-)]}{1 - \epsilon a - a/[1 + 1/kI(z_-)]} \cdot (11)$$

For $a = 0$, the denominator of Eq. 11 is 1.0; the numerator of Eq. 11 is identical to the first two terms of the right-hand side of Eq. 9. Therefore, for $a = 0$, Eq. 11 is equivalent to Eq. 10, which used only the first two right-side terms of Eq. 9. More generally, Eq. 11 may be written in terms of Eq. 10 as

$$\Delta\ell = \frac{c_1 + c_2\ell}{1 - c_3(\ell)} \quad (12)$$

where $c_3(\ell) = \epsilon a + akI(z_-)/[1 + kI(z_-)]$. For extreme values of ℓ, the limits of $c_3(\ell)$ are

$$\lim_{\ell\to\infty} c_3(\ell) = a(1 + \epsilon) \quad (13a)$$

$$\lim_{\ell\to 0} c_3(\ell) = a\epsilon. \quad (13b)$$

This means that $\Delta\ell$'s computed from Eq. 10 (the simplified solution of Eq. 6) must be increased by a correction factor of $(1 - a\epsilon)^{-1}$ at low background luminances and a correction factor of $[1 - a(1 + \epsilon)]^{-1}$ at high luminances to satisfy the exact solution. These factors depend on the area of T. For test stimuli that are points, the correction factors differ insignificantly from 1.0, and Eq. 10 remains valid. For test stimuli of large area, the correction factor becomes about 2 at high luminance and, therefore, is significant. Details of the definition and calculation of the area factors are given below.

The effect of test stimulus area on $\Delta\ell$ vs ℓ. (1) Point-stimulus. In the expansion of Eq. 8 into Eq. 11, a term containing the ratio of inhibition to excitation occurred repeatedly, and it was convenient to designate it by the symbol a. In fact, a is the ratio of two pure numbers

$$a = \left[\frac{\Delta I}{I(z_-)}\right] \cdot \left[\frac{\Delta H}{H(z_-)}\right]^{-1} = \beta/\alpha. \quad 4)$$

In words, a is the ratio of the proportional increase in inhibition β to the proportional increase in excitation α. It reflects the amount by which T increases inhibition relative to the amount by which T increases excitation. For the case of M and B fields that uniformly cover the entire area of interest, a reduces to $\sigma^{-2}\,\Delta I/\Delta H$, the form used in Eq. 11. Finding the effect of test area on threshold $\Delta\ell$ involves first calculating a. The first case to be considered is: T = a point.

A point stimulus at (0,0) is defined as $\ell_T(x,y) = \delta(0,0)\ell_T$. The factor ΔH is

$$\Delta H = \ell_T(x,y) * w_H(x_+ - x, y_+ - y) - \ell_T(x,y) * w_H(x_- - x, y_- - y) \quad (15)$$

Maximum excitation occurs at $z_+ = 0,0$; minimum excitation occurs Δx away, at $z_- = (\Delta x, 0)$. Writing w_H as a function of distance Δx and letting E_T denote the total energy in the test gives

$$\Delta H = E_T[w_H(0) - w_H(\Delta x)] = E_T[1 - w_H(\Delta x)],$$

and

$$\Delta I = E_T[w_I(0) - w_I(\Delta x)] = E_T[1 - w_I(\Delta x)].$$

The total energy contributed to H at z_+ (and at z_-) by a uniform masking field is simply the effective area W_H of w_H times ℓ_M: $H = \ell_M W_H$. If w_I is of the same form as w_H, but larger by a factor of $\sigma[w_I(x) = w_H(\sigma x)]$ then the effective area W_I of w_I is $\sigma^2 W_H$, and $I = \sigma^2 \ell_M W_H$. Putting these terms together in Eq. 14 gives (for a point stimulus):

$$a = \left\{\frac{E_T[1 \quad w_I(\Delta x)]}{\sigma^2 \ell_M W_H}\right\}\left\{\frac{\ell_m W_H}{E_T[1 - w_H(\Delta x)]}\right\}$$

$$= \frac{1}{\sigma^2}\left[\frac{1 - w_I(\Delta x)}{1 - w_H(\Delta x)}\right]$$

Taking

$$w_H(x) = e^{-x^2/2},\ w_I(x) = e^{-x^2/(2\sigma^2)},$$

$$\sigma = 3 \text{ and } \Delta x \approx 2.80$$

gives[5]

$$a = \frac{1}{9}(0.351)/(0.980) = 0.038.$$

The effect of test stimulus area on $\Delta\ell$ vs ℓ. (2) Half-field stimulus. For a half-field $\ell_T(x,y) = u(x)\ell_T$, where $u(x)$ is a unit step at the point zero $[u(x) = 0, x < 0; u(x) = 1, x \geq 0]$. When ℓ_T is very small, the maximum and minimum of G are symmetrical around the boundary, $z_+ = -z_-$. Choosing w_H and w_I as above gives[6] $z_+ = \Delta x/2 \approx 1.57$. Substituting $\ell_T(x,y)$ in Eq. 15 gives:

$$\Delta H = \ell_T\left[\int_{\Delta x/2}^{\infty} w_H(x)dx \right.$$

$$\left. \int_{+\Delta x/2}^{\infty} w_H(x)dx\right]$$

$$= \ell_T \int_{\Delta x/2}^{\Delta x/2} w_H(x)dx. \qquad (16)$$

The calculation of ΔI is the same as Eq. 16 with w_I substituted for w_H.

To simplify Eq. 16, it is convenient to define the fractional overlap of T with w_H at z_+ as $p(z_+) = \ell_T(x,y) * w_H(x_+ - x, y_+ - y)/(\ell_T W_H)$. When $\ell_T(x,y)$ covers all of w_H at z_+ then $p(z_+) = 1$. The range of p is $0 \leq p \leq 1$. By defining $p(z_-)$ similarly (with z_- substituted for z_+), and defining $\Delta p = p(z_+) \quad p(z_-)$, ΔH may be written as $\ell_T W_H \Delta p$. Analagous fractional overlaps of ℓ_T with w_I may be defined as $q(z_+)$, $q(z_-)$, and Δq. Substituting these into Eq. 14 gives

$$a = \left(\frac{\ell_T \cdot \sigma^2 W_H \cdot \Delta q}{\ell_M \cdot \sigma^2 W_H}\right)\left(\frac{\ell_M W_H}{\ell_T W_H \Delta p}\right) = \frac{\Delta q}{\Delta p}. \qquad (17)$$

Equation 17 was derived without any specific assumptions about geometry and, therefore, holds for all test stimuli. Note that the area-factor σ^2 appears wherever W_I appears, i.e., in both the numerator and denominator of $\Delta I/I$, and therefore σ^2 is cancelled in Eq. 17.

Using the same normal functions for w_H and w_I in the case of the half-field as in the point test gives, for the half-field: $p(z_+) \approx 0.94$, $p(z_-) \approx 0.06$, $q(z_+) \approx 0.70$, $q(z_-) \approx 0.30$, so that $a = \Delta q/\Delta p = (0.40)/(0.88) \approx 0.45$. This value of a cannot be ignored in Eqs. 11 and 12, which shall now be reconsidered.

Effect of test stimulus area on $\Delta\ell$ vs ℓ. (3) General form. According to Eqs. 13a and 13b, the area correction is smallest at low retinal illuminances, when it is $1 - a\epsilon$. The threshold criterion ϵ is assumed to be small, that is, $\epsilon \leq 0.1$. Using the values of a that were calculated above, we see that $a\epsilon < 0.045$ even for large-area tests; the area correction always is negligible at zero background illuminance. At high illuminances, the correction factor is $[1 - a(1 + \epsilon)]^{-1}$. For point tests ($a \approx 0.038$), the area correction differs insignificantly from 1.0. For large-area tests, $a = 0.45$; assuming $\epsilon = 0.1$ gives $[1 \quad a(1 + \epsilon)]^{-1} = (0.505)^{-1} \approx 2.0$, i.e., an area correction of about 2 for $\Delta\ell$ vs ℓ.

For the case of an incremental test stimulus of illuminance ℓ_T on a large uniform background ℓ it is possible to write an almost-explicit exact form of Eq. 11 to define the threshold value $\Delta\ell$ of ℓ_T. Using Δp as defined above, ignoring time effects, assuming large, uniform M and B fields gives

$$\Delta\ell = \frac{\left[\frac{\epsilon}{k\sigma^2 W_H \Delta p}\right] + \left[\frac{\epsilon}{\Delta p}\right]\ell}{\left[1 \quad \epsilon a - \frac{a}{1 + 1/(k\sigma^2 W_H \ell)}\right]} \qquad (18)$$

The area parameter a must be calculated separately for each test stimulus configuration. It must also be calculated separately for each background illuminance ℓ because a depends on z_- and z_+, and z_- and z_+ vary somewhat with the values of ℓ and of $\Delta\ell$. Fortunately, using the same value of a for all ℓ_T, ℓ_B, introduces only a minor error into Eq. 18.

The parameter Δp, which describes the difference in overlap of T with w_H at z_+ and at z_-, reaches values near 1.0 for large test fields (0.884 of a possible 1.0 for the w_H and w_I assumed in the examples) and Δp goes to zero as the area of T goes to zero. This means that the main effect of test area on threshold occurs in the calculation of Δp; the effect of test area via a is only of second-order importance.

The meaning of ϵ. For a large uniform field B, the asymptotic value of G as $\ell_B \to \infty$ is $G = H/kI = 1/(k\sigma^2)$. Assuming that the threshold of detection ΔG is some fractional value ϵ of the limiting value of G gives: ΔG (threshold) $= \epsilon/(k\sigma^2)$, or $\epsilon = k\sigma^2 \Delta G$. This is the definition of ϵ given earlier. It also is presumed that ϵ is small, i.e., $\epsilon \leq 0.1$.

Equation 18 indicates another reason for defining the threshold ϵ as above. Together with Δp, ϵ gives the main factor in the slope of $\Delta\ell$ vs ℓ, $d\ell_T/d\ell_B$, the Weber fraction. When Δp is 1.0 (T overlaps perfectly with w_H at z_+ and not at all with w_H at z_-) ϵ is simply the fractional amount by which ℓ_T must increase the flux being received by w_H at z_+ in order to bring the increase to threshold. When $\Delta p < 1$, the Weber fraction is larger than ϵ because T stimulates w_H at z_+ imperfectly $[p(z_+) < 1]$ and/or because it stimulates w_H at z_- $[p(z_-) > 0]$. The effect of T on inhibition (w_I) is smaller, and it is subsumed in the area term a of the denominator of Eq. 18.

Critical area. When ℓ is large, Eq. 18 reduces to a simpler form,

$$\lim_{\ell \to \infty} \Delta\ell = \frac{(\epsilon/\Delta p)\ell}{1 - a(1 + \epsilon)} = \frac{\epsilon\ell}{\Delta p \quad (1 + \epsilon)\Delta q} \qquad (19)$$

From Eq. 19, it is possible to predict the observed "critical area" of the feedforward field; i.e., the intersection of two asymptotes in a graph of the energy of a threshold test stimulus vs the area of the stimulus. The equation of the small-area asymptote is $A \cdot \Delta\ell = E_o$ (Ricco's law), where A is the area of the test stimulus, and E_o is the minimum energy for detection. The equation of the large-area asymptote is $A_c \cdot \Delta\ell = E_o$, where A_c is the critical area. To solve for critical area, these two quantities are set equal, namely

$$A \cdot \Delta\ell_o = A_c \Delta\ell_1$$

giving

$$A_c = A \cdot \Delta\ell_o/\Delta\ell_1. \qquad (19a)$$

The subscript o denotes a small test field (i.e., a point) and the subscript 1 denotes a large test field (i.e., a half-field). To calculate A_c, we substitute Eq. 19 into Eq. 19a, giving

$$A_c = A\left\{\frac{\epsilon\ell}{\Delta p_o[1 - a_o(1 + \epsilon)]}\right\} \cdot \left\{\frac{\epsilon\ell}{\Delta p_1[1 - a_1(1 + \epsilon)]}\right\}^{-1}$$

$$A_c = \frac{A\Delta p_1[1 - a_1(1 + \epsilon)]}{\Delta p_o[1 - a_o(1 + \epsilon)]} \qquad (19b)$$

Equation 19b is valid for large values of ℓ. To reduce Eq. 19b further, it is necessary

to calculate the Δp's and the a's. This has already been done for the case of point T's ($\Delta p_o, a_o$) and for half-field T's ($\Delta p_1, a_1$). For T = a point, $\Delta p_o = [A/W_H][1 - w_H(\Delta x)] \approx 0.98\, A/W_H$, and $a_o = 0.038$. For T = a half-field, $p_1 \approx 0.88$, independent of A, and $a_1 = 0.45$. Substituting the algebraic quantities into Eq. 19b yields

$$A_c/W_H = \frac{\Delta p_1 [1 - a_1(1+\epsilon)]}{[1 - w_H(\Delta x)][1 - a_o(1+\epsilon)]}$$

$$\approx \Delta p_1 (1 - a_1) \qquad (19c)$$

The last expression of Eq. 19c is a convenient, rough approximation. Substituting the numerical values obtained for the case where w_H, w_I are normal distributions, and $\sigma_I/\sigma_H = 3$, gives

$$\lim_{\ell \to \infty} A_c/W_H \approx 0.47.$$

A similar calculation of the critical area of the dark-adapted model shows it to differ negligibly from W_H.

General solution. The values of H, I, and G are defined by Eqs. 1, 2, and 3a, even when a closed-form solution does not exist. By using summations to approximate integrals, G may be calculated for any explicitly defined ℓ_T, ℓ_M, ℓ_B. The author is indebted to Mrs. J. T. Budiansky for a computer program to calculate G(x,y) for the special case of T = a disk stimulus of radius r, B = a uniform background of illuminance ℓ, and the weighting functions w_H, w_I proportional to normal distributions (Fig. 3). The results of these computations are illustrated later, in Section III.

Transient effects. In vision, transient responses generally are greater than steady-state responses, and test stimuli are detected by their transient response. In the model, transient effects are subsumed under the temporal weighting functions, v_H, v_I, which are not known exactly. In this section, some conclusions are derived from very general considerations about v_H, v_I.

In considering temporal effects, the concern here is primarily with the form of the $\Delta\ell$ vs ℓ function rather than with the particular values of $\Delta\ell$. E.g., the basic form of the $\Delta\ell$ vs ℓ function is $\Delta\ell = c_1 + c_2\,\ell$. The temporal weighting functions affect c_1 and c_2. On a graph of $\Delta\ell$ vs ℓ, changes in c_1 and c_2 merely shift the curve vertically and/or change the scale of the horizontal coordinate; they do not change the form of the curve. When area effects are significant, $\Delta\ell$ is given by Eq. 12, which contains a factor of $[1 - c_3(\ell)]^{-1}$. The factor $c_3(\ell)$ depends on ℓ, producing changes in the shape of $\Delta\ell$ vs ℓ, which cannot be compensated by linear transformations of the coordinates. Because $H(z_+)$, $H(z_-)$, $I(z_+)$, $I(z_-)$ occur in $c_3(\ell)$ only in terms of ℓ directly and in terms of the area factor a, the problem reduces to that of finding the effect of temporal stimulus variations upon a.

In Eq. 14, a was defined as the ratio of $\beta = \Delta I/I(z_-)$ to $\alpha = \Delta H/H(z_-)$. Visual responsiveness to transients means that test stimulus illuminance ℓ_T contributes more effectively to ΔH and to ΔI than background illuminance ℓ_B contributes to $H(z_-)$ and $I(z_-)$. Let η_H be the factor expressing the increased effectiveness of a transient input to ΔH relative to that of a steady-state input; i.e., we use $\eta_H \ell_T$ to calculate ΔH rather than ℓ_T. Let η_I be the factor expressing the increased effectiveness of a transient input to ΔI relative to that of a steady-state input. Let $\eta = \eta_I/\eta_H$. By substituting η_H and η_I into Eq. 14, the revised area factor a' is seen to be $a' = \eta a$.

The effect of temporal variations on the form of $\Delta\ell$ vs ℓ is thus given by a single parameter, η. When T and B have the same temporal waveshape [i.e., $\ell_T(t) = c\ell_B(t)$], then by definition $\eta = 1.0$. When T is a step function, superimposed on a steady background B, then the effective value of η will vary with time. Suppose inhibition grows more slowly than excitation [for example, $v_I(t) = v_H(mt)$, $m > 1$]. Then η becomes a function $\eta(t)$ of the time t after onset of the step. Initially $\eta(t) < 1$, finally $\eta(t) = 1$, and for reasonable assumptions about $v_H(t)$, $v_I(t)$, $\eta(t)$ will overshoot the value 1.0 for some intermediate value of t.[7]

The variation of η with time means that the spatial response characteristics vary with the duration of a test stimulus. In considering psychophysical data, a simplifying assumption will be made; namely, at a moment of detection t_d, the model can be represented by the currently effective value of η, $\eta(t_d)$.

IId. Sine-Wave Contrast Thresholds

A sine-wave contrast stimulus is defined by $\ell(x,y)$,

$$\ell(x,y) = \ell(1 + m \cos \omega x).$$

$\ell(x,y)$ is a one-dimensional stimulus in the sense that retinal illuminance varies only in the x-dimension. For a given threshold criterion ΔG of the feedforward model, and for a given average illuminance ℓ, the modulation threshold m is a function of ω. The function depends on ℓ and on ΔG. By assuming, as before, that

$$w_H(x) = e^{-x^2/2\sigma_H^2}) \text{ and } w_I(x) = e^{-x^2/(2\sigma_I^2)}$$

an exact solution is derived to relate all the factors.

When a sine stimulus $\ell(x,y)$ is the input to the feedforward system, the output is:

$$G(x,y) = \frac{\ell(x,y) * w_H(x)}{1 + k\ell(x,y) * w_I(x)}$$

$$= \frac{\ell(1 + m \cos * w_H)}{(1 + k\ell) + mk\ell \cos * w_I} \qquad (20)$$

To find the threshold it is necessary only to determine $\Delta G = \max G(x) - \min G(x)$. The algebra is simplified by disposing early of unneeded terms. Equation 20 is of the form $(c_1 + v_1)/(c_2 + v_2)$ where c and v denote constant and variable terms (as a function of x). Furthermore, the maximum of G occurs when x = 0, the minimum occurs at $x = \pi/\omega$, so that $v_1(x_+) = -v_1(x_-)$ and $v_2(x_+) = -v_2(x_-)$. Thus ΔG is of the form

$$\Delta G = c\left[\frac{1+a}{1+b}\right] - c\left[\frac{1-a}{1-b}\right] = 2c\frac{a-b}{1-b^2} \qquad (21)$$

Before substituting for a, b, and c in Eq. 21 it is desirable to change scale factors slightly and to define $w'_H = (2\pi)^{-\frac{1}{2}} \sigma_H^{-1} w_H$ and $w'_H = (2\pi)^{-\frac{1}{2}}\sigma_I^{-1} w_I$, so that $w'_H(x)$ and $w'_I(x)$ are probability distribution functions. By defining $\Delta G' = (2\pi)^{-\frac{1}{2}}\sigma_H^{-1}\Delta G$ and $k' = (2\pi)^{+\frac{1}{2}}\sigma_I k$ the primed system becomes equivalent to the unprimed system. In Eq. 20, and in all equations derived from it, changing from unprimed to primed w_H, w_I, ΔG, and k preserves the equalities.

Considering Eq. 20 to be of the form $c(1 + a)/(1 + b)$ gives the following identifications of a, b, c:

$$a = m \cos * w'_H = m \exp\left[-\tfrac{1}{2} n^2 \sigma_H^2\right];$$

$$b = m\gamma \cos * w'_I = m \exp\left[-\tfrac{1}{2} n^2 \sigma_I^2\right];$$

$$c = \gamma/k';$$

$$\gamma = k\ell/(1 + k\ell).$$

Substituting these values into Eq. 22 gives

$$\Delta G' = m(2\gamma/k) \exp\left[-\tfrac{1}{2} n^2 \sigma_H^2\right]$$

$$\left[1 - \gamma \exp\left[-\tfrac{1}{2}(\sigma_I^2 - \sigma_H^2)\right]\right]$$

$$/\left[1 - m^2\gamma^2 \exp\left(-n^2\sigma_I^2\right)\right]. \qquad (23)$$

Equation 23 is of the form

$$A = \frac{m}{1 - m^2 B}, \qquad (24)$$

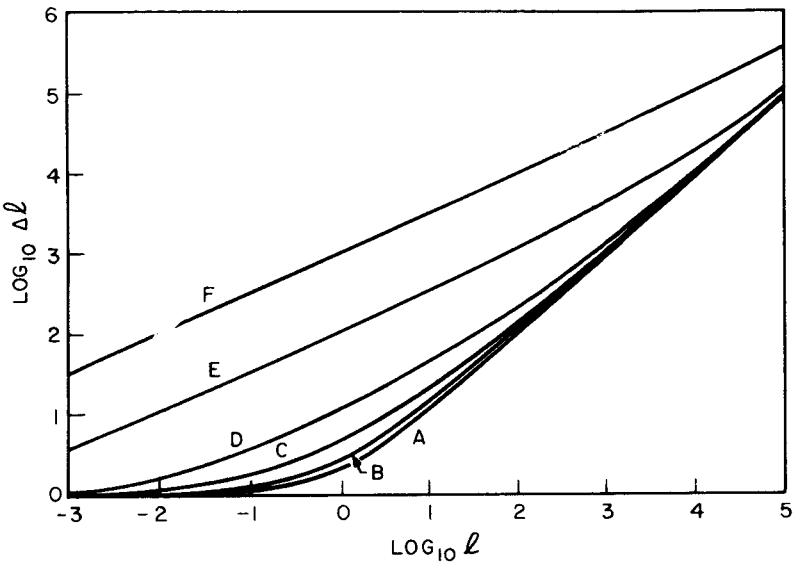

Fig. 6. Theoretical log $\Delta\ell$ vs log ℓ (test threshold vs adapting background) functions. (a) Pure feedforward forward model; (b-f) models with feedback plus feedforward (see Fig. 7). The ratio r of feedback to feedforward for each of the curves is (a) r = 0, (b) r = 1, (c) r = 10, (d) r = 10^2, (e) r = 10^4, (f) r = 10^6.

where

$$A = k'\Delta G'(2\gamma)^{-1} \exp(\tfrac{1}{2} n^2 \sigma_H^2) / [1 - \gamma \exp[-\tfrac{1}{2} n^2 (\sigma_I^2 - \sigma_H^2)]]$$

and

$$B = \gamma^2 \exp[-n^2 \sigma_I^2].$$

Equation 24 is a quadratic equation in m; the positive root is

$$m = (2AB)^{-1}\left[\sqrt{1 + 4A^2B} - 1\right], \quad (25)$$

where A and B are defined above. Equation 25 is an exact expression for the input modulation depth m needed to produce a response of $\Delta G'$ in the output of the feedforward system. However, Eq. 25 does not take into account any prior feedback adaptation.

IIe. Prior Feedback Adaptation

Incremental thresholds. Let $\Delta\ell$ be the incremental threshold for a feedforward system and $\Delta\ell'$ be the incremental threshold for the same system preceded by n stages of shunting feedback. If we take the gain of each stage as unity, it can be shown from Eq. 4 that in the steady-state

$$\frac{\partial}{\partial \ell}\Delta\ell' = \left[\frac{Y + 1}{Y + \frac{1}{n+1}}\right]\Delta G \quad (26)$$

where Y is the output of the n^{th} stage. For large ℓ, Y also is large, i.e., $Y \rightarrow \ell 1/(n+1)$, so that $\Delta\ell' \rightarrow \Delta\ell' + 0(\Delta\ell)$, where $0(\Delta\ell)$ represents a quantity that becomes negligible with respect to $\Delta\ell$ as $\ell \rightarrow \infty$. Similarly, as $\ell \rightarrow 0$, $\Delta\ell' \rightarrow (n+1)\Delta\ell + C \rightarrow C'$: At high and at low levels of input, therefore, the behavior of the feedback-plus-feedforward system and of simple feedforward system is identical. In fact, when n is small, the two kinds of systems are quite similar at all levels of input; n merely determines the gradualness of the transition from the constant response characteristic at small inputs to the proportional response characteristic at large inputs.

So far, it has been assumed that the feedback and the feedforward components were equally weighted in the sense that the feedback signal becomes effective in the same range of illuminances as the feedforward signal. This assumption is unimportant when the input is very small or very large. For example, when the input is sufficiently small, both the feedback and the feedforward stage behave like simple transducers; their output is proportional to the input. Feedback and feedforward become significant only when intermediate values of ℓ are attained.

If the feedback stage became effective at a much lower input level than the feedforward stage, then the overall response of the system would go through three distinct phases: (1) for very small ℓ, $\Delta\ell \approx c_1$, (2) for intermediate ℓ, $\Delta\ell \approx \ell 1/(n+1)$, (3) for large ℓ, $\Delta\ell \approx c_2\ell$. Significant feedback at small ℓ increases the size of the transition zone between constant $\Delta\ell$ (small ℓ) and proportional $\Delta\ell$ (large ℓ) by interposing a zone of slope $\approx 1/(n+1)$. For the case of n = 1 (one stage of feedback adaptation) an explicit expression for $\Delta\ell$ is derived from Eqs. 3a and 4, in terms of the amounts of feedback k_1 and feedforward k_2; namely,

$$\Delta\ell = (k_2 x + \sqrt{k_1 x + \tfrac{1}{4}} + \tfrac{1}{2})\Delta G. \quad (27)$$

III. PROPERTIES OF THE MODEL

Light Adaptation: Weber Law and Square-Root Law

Change in $\Delta\ell$ with ℓ. Consider detection of contrast at the boundary of a small incremental test field of retinal illuminance $\Delta\ell$ superimposed on a background field of retinal illuminance ℓ. To produce a small, constant output ΔG in the feedforward neural field, Eq. 6 showed that $\Delta\ell = c_1 + c_2\ell$, where the values of c_1 and c_2 depend on the spatial geometry and temporal waveforms of the test and background fields. This predicted relation between $\Delta\ell$ and ℓ is illustrated in Curve A of Fig. 6. The horizontal asymptote represents the dark-adapted threshold c_1; the diagonal of slope = 1 (the high-illuminance asymptote) represents a Weber law with intercept c_2. That is,

$$\lim_{\ell \rightarrow \infty} \Delta\ell/\ell = c_2.$$

The dimensionless constant c_2 is called the Weber constant for contrast detection.

Figure 6 also illustrates five cases in which shunting-feedback adaptation precedes the feedforward stage. The amounts of feedback and feedforward are given by r and 1, respectively (Fig. 7). In the case of pure feedforward, r is 0. When feedback and feedforward control signals are evenly matched (r = 1, Curve B in Fig. 6) the $\Delta\ell$ vs ℓ function is barely discriminable from the case of pure feedforward. As feedback becomes relatively more significant (r = 10, 10^2, 10^4, 10^6, respectively), the predicted thresholds $\Delta\ell$ follow a square-root law (slope = ½) for a bigger and bigger region between the two unchanging asymptotes of slope 0 and 1. No matter how large the

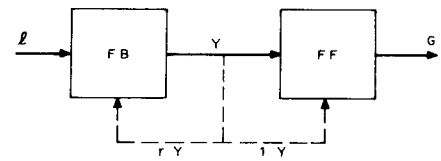

Fig. 7. Block diagram of the model. FB is feedback component; FF is feedforward component; ℓ is input retinal illuminance; G is neural output; r is the relative amount of signal fed back (FB/FF). The detector is not indicated.

ratio of feedback to feedforward signal may be, the same asymptotes ultimately are achieved!

Figure 6 illustrates how the nature of the transition from dark-adapted behavior (horizontal line) to light-adapted behavior (line of slope 1) can be determined by the ratio of feedback to feedforward signals. For example, a line of slope ½ over a horizontal range of about 3 logs (Curve C) occurs when $r = 10^2$. The threshold relation of $\Delta\ell \propto \ell^{1/2}$ is a square-root law. It is commonly observed at low levels of ℓ, particularly in peripheral viewing, i.e., in rod vision. According to the model, the difference between threshold functions observed in different regions of the retina is explained by a regional variation in the parameter r.

In terms of neural mechanism, a large value of r (in the peripheral retina) requires that feedback signals be mediated by special cells—possibly horizontal cells. This is because the feedback is effective even when only one quantum per hundreds of rods is being absorbed from ℓ, so that the probability of a quantum from $\Delta\ell$ striking a rod that recently received a quantum from ℓ is small. Adaptation of a rod is determined by its environment (e.g., the rod pool, Rushton, 1963). Although the same also may be true for foveal vision (i.e., cones), it cannot be proved on a priori grounds, because cones respond only at higher levels of retinal illumination (e.g., several quanta/cone; Brindley, 1960, p. 187ff) and therefore the feedback loop theoretically could be within a single receptor.

Shunting feedback: An evolutionary adaptation to quantum noise. At low levels of illumination, visual discrimination approaches the limits set by photochemical absorption and the quantum variability of the stimulus (Hecht et al, 1942; DeVries, 1943; Pirenne, 1962; Barlow, 1962b). That is, the performance of the eye approaches the performance of a perfect detector which absorbs the same number of quanta as does the eye (Barlow, 1962a). At high retinal illuminances, the eye is not nearly so efficient.

The measure of quantal (Poisson-distributed) stimulus noise is the noise's standard deviation, i.e., its root-mean-square (RMS) value. If the limit of visual discrimination were determined by quantum noise of the stimulus, then the detection threshold $\Delta\ell$ would increase as $\ell^{1/2}$, because the RMS value of quantum noise increases as $\ell^{1/2}$. If there were no visual adaptation, test signals $\Delta\ell$ of equal detectability would then produce vastly different visual responses at the point of detection; i.e., $\Delta\ell$ signals from high-ℓ backgrounds would produce much larger signals than equally detectable $\Delta\ell$ signals on low-ℓ backgrounds. To utilize the stimulus information efficiently, the O would have to adjust his detection criterion delicately to each background level ℓ. In fact, human Os are notoriously unable to make accurate judgments of background retinal illuminance, which poses the problem of how Os could adjust a criterion to depend on something (ℓ_B) that they cannot estimate. On the other hand, humans are able to make stable detection judgments of $\Delta\ell$, even with little practice. These difficulties are resolved naturally by the shunting feedback component.

Shunting feedback produces an output that is asymptotically proportional to $\ell^{1/2}$. Thus, intensity discrimination would be impaired because of the reduced dynamic range of the output. On the other hand, small threshold increments of $\alpha \cdot \ell^{1/2}$ would produce output increments proportional to α and therefore would be equally detectable. The shunting feedback system, which reduces the input by a factor of $\ell^{1/2}$, is the ideal match to a stimulus in which RMS noise increases by a factor of $\ell^{1/2}$. In the output of a shunting feedback system, the RMS value of quantal noise in ℓ remains constant, independent of ℓ. Thus, as ℓ varies, increments $\Delta\ell$ of constant statistical significance produce constant incremental outputs, independent of ℓ. With one stage of shunting feedback, the detection rules can be independent of ℓ—an enormous simplification of the detection problem.

The second remarkable property of shunting feedback is that it does not alter the desirable properties, at high input luminances, of a subsequent feedforward component. The Weber-law characteristics of the system result from the feedforward component; these characteristics are left entirely intact by shunting feedback.

Additivity of illuminance components. Consider a stimulus composed of several components; for example, of a brief masking flash and of a steady background (Fig. 4). The threshold luminance measured in the presence of each component stimulus, individually, is $\Delta\ell_1$, $\Delta\ell_2$, $\cdots$, $\Delta\ell_n$, etc., and the threshold in the presence of all components together is $\Delta\ell_{1\,+\,2\,\cdots\,+\,n}$. Then the model (Eq. 10) predicts that

$$\max(\Delta\ell_i, \cdots) \leqslant \Delta\ell_{1\,+\,2\cdots+n} \leqslant \Sigma\,\Delta\ell_i . \qquad (28)$$

The left equality holds for low retinal illuminances (where the $\Delta\ell$ vs ℓ function is flat) and the right equality holds at high illuminances (where $\Delta\ell$ is proportional to ℓ). For the right-hand equality to hold, each compoment ℓ_i must already be in the Weber-law range, and the $\Delta\ell$'s must be taken in the same direction (e.g., all increments or all decrements).

The right-hand equality has not been widely tested. Sperling (1965) found it to hold over a wide range of masking flashes superimposed on backgrounds of varying luminances (cf. Fig. 4), but his range of backgrounds was limited.

Many theorists have proposed that the visual system behaves linearly for small signals. Small-signal linearity is a trivial consequence of any realizable continuous system (i.e., also having a continuous derivative). The problem with small-signal linearity is that the linear range may be very small. The significance of Eq. 28 is that it predicts superposition to hold for large signals; the larger the background input signals, the better it holds. Considering how nonlinear the elements of the model are, it is remarkable that additivity should ever obtain, all the more remarkable that the model should become a linear system with large masking signals.

Receptive Fields

Consider the output of the model G(x,y) at a particular point x,y. Let the input be a point of light $\Delta\ell(x',y')$ superimposed on a steady uniform background ℓ. It was assumed (Section II) that the excitatory w_H and inhibitory w_I weighting functions were radially symmetrical, so that G is a function only of $\Delta r = \sqrt{(x-x')^2 + (y-y')^2}$, the distance of spot $\Delta\ell$ from x,y. When $y' = y$, then $\Delta r = \Delta x$.

To calculate G(r), specific distributions must be given for $w_H(r)$ and $w_I(r)$. In the examples of this section, it is assumed that w_H and w_I are proportional to two-dimensional normal distributions, with the standard deviation σ_I of w_I equal to $3\sigma_H$ (Fig. 3). In terms of area, the effective area W_I of w_I is nine times the effective area W_H of w_H: $W_I/W_H = (\sigma_I/\sigma_H)^2$.

The shape and magnitude of the response G(r) to $\Delta\ell$ depend on both ℓ and $\Delta\ell$. Three values of ℓ are considered: $\ell = 0$, 1/18, and 10^3, representing darkness, a dim background, and an asymptotically-intense background. The input value of $\Delta\ell$ is chosen arbitrarily so that—in the absence of self-inhibition—it would produce a maximum response ΔG of magnitude equal to the asymptotic steady-state value of G in the light (e.g., so that

$$\max \Delta G = \lim_{\ell\to\infty} G(\ell_B) = 1/k\sigma^2).$$

Because the self-inhibiting effect of $\Delta\ell$ on itself is neglected in choosing the value of $\Delta\ell$, the actual value of max ΔG is somewhat less than $1/k\sigma^2$.

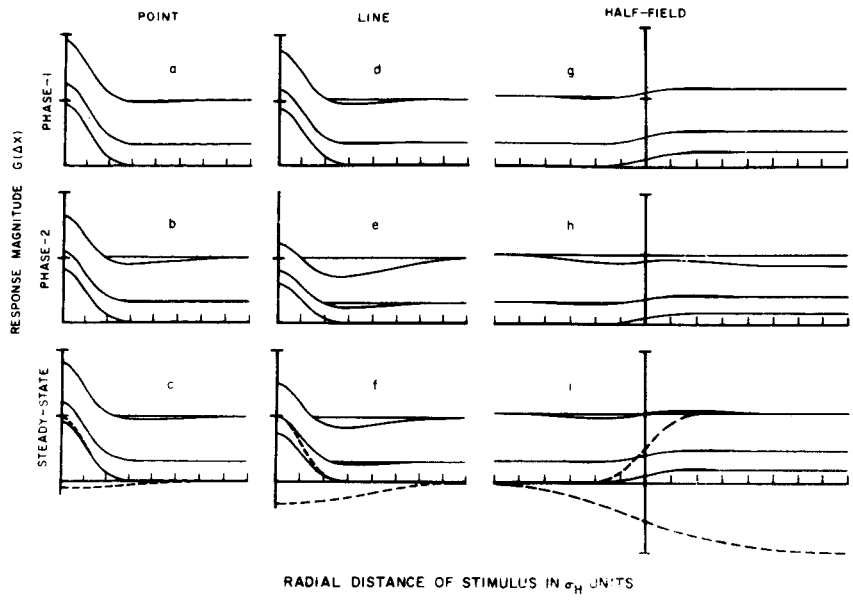

Fig. 8. Output of the feedback component of the model at one point (x,y) as the location (x + Δx,y) of a test stimulus is varied. Horizontal scale markings indicate radial distances of 1.0 σ_H. Sections a, b, c show responses to a point stimulus; Sections d, e, f show responses to a line stimulus; Sections g, h, i show responses to a half-field stimulus. The retinal illuminance Δℓ of the line and of the half-field stimuli is such that when they are optically superimposed on w_H they provide w_H with exactly as much light as the corresponding point stimuli. Responses are illustrated at three different times after onset of Δℓ: Phase 1 (a,d,g) is peak of initial transient, η = 2; Phase 2 (b,e,h) transient rebound, η = 0.5; steady-state (c,f,i) η = 1.0. The three different curves in each section represent responses G(Δx,ℓ,Δℓ) calculated at three different levels of background ℓ: darkness (lowest curves), asymptotically intense ℓ (highest curves), and an intermediate level of ℓ (middle curves). The dashed curves in Sections c, f, i indicate test-induced excitation ΔH and inhibition ΔI separately (−ΔI is illustrated). The illuminance Δℓ of the test stimuli was so chosen that the maximum value of ΔH is the same for all 27 curves; within a section the entire curves for ΔH and ΔI are the same at all three levels of background ℓ. See text for more details.

Consider first Fig. 8c. It illustrates the steady-state response G(x,y) to a point Δℓ as a function of Δr, the distance of Δℓ from x,y. In darkness, the response G has no undershoot. At high backgrounds, the response G has a slight undershoot. The curves illustrating the response G(x,y) to Δℓ(x + Δr,y) are analogous to the usual physiological representation of neural responses, e.g., of the firing rate of a retinal ganglion cell as a function of the retinal location of a stimulating spot of light. The reasons for identifying the feedforward stage with retinal ganglion cells are set forth by Sperling and Sondhi (1968, p. 1139). The w_H function corresponds to the center of the receptive field and the w_I function to the surround.

To represent visual processes more accurately, temporal factors need to be taken into account. Using a similar model (feedback plus feedforward), Sperling and Sondhi (1968) showed that its temporal response to an impulse of light was monophasic at low background luminances and biphasic at high luminances. The surround portion contributed its response more slowly than the center and thus seemed to be delayed. Step inputs to the feedforward system may be considered to induce three recognizable phases. In Phase 1, the excitatory input H is at its maximum overshoot (e.g., two times the steady-state value) and the inhibitory component has acquired its steady-state value. In Phase 2, the excitatory component has subsided to its steady-state value, but the inhibitory component is overshooting to two times its steady-state value. Phase 3 is the steady-state, described above.

Figures 8a and 8b illustrate the point-response of the model in Phase 1 and Phase 2, respectively, with these values of overshoot and undershoot. The actual amount by which transient responses exceed steady-state responses is immaterial to the shape of the spatial response characteristics illustrated in Fig. 8. The various temporal effects on spatial response characteristics were shown in Section IIc to be determined by a single parameter η which is a ratio of two ratios, β/α: β is the ratio of effectiveness of transient (test) inhibition to steady-state (background) inhibition, and α is the ratio of effectiveness of transient (test) excitation to steady-state (background) excitation. For Phase 1 and Phase 2, η is assumed to be 0.5 and 2.0, respectively; in the steady state, η is 1.0 by definition.

In the usual type of physiological recording, the maximum of response is noted without regard for the precise time of occurrence of the maximum. Therefore, the appropriate responses to compare to physiological recordings are the maxima of Phase 1 combined with the minima of Phase 2. However, the main features of the comparison are independent of these details. First, they show that in the dark the receptive field has a response that is characteristic only of the center. With increasing light adaptation, the receptive field shows, more and more, the effects of an antagonistic surround. The light-adapted receptive field contains areas that had responded with the center characteristic in the dark, but which respond with the surround characteristic in the light. Thus the effective diameter of the center of the receptive field shrinks with light adaptation. These effects of light adaptation occur without any change of connectivity of the model; they simply reflect the nonlinear, intensity-dependent mode of response of the feedforward component.

Comparison of the Phase 1 and Phase 2 responses shows that: for small values of Δr, both Phase 1 and Phase 2 responses are positive; for intermediate values of Δr, Phase 1 responses are positive and Phase 2 responses are negative; for large values of Δr, both Phase 1 and Phase 2 responses are negative. The areas defined by these values of Δr correspond to the *on-center, on-off surround,* and *pure-off surround* of an *on-center* receptive field. These areas are similarly defined for off-center receptive fields, except that the sign of the model's response must be reversed.

Impulse response (point-spread function). The curves of Figs. 8a, 8b, and 8c give the output G(x,y) at a single point x,y, as the location x + Δx,y of a stimulating spot is varied. A little reflection will convince the reader that when the spot is stationary at x,y and the output is measured at different locations, x + Δx,y, the resulting curves describing G(x + Δx,y) are identical to those shown in Fig. 8. This symmetry between input and output exists whenever the properties of the

transformation remain homotropic (independent of absolute location). It follows that the curves of Figs. 8a, 8b, and 8c describe the spatial impulse response of the model (i.e., its response to an impulse in space, a point). In fact, Fig. 8 illustrates only half of the spatial impulse response; the response is mirror-symmetrical around its maximum. The full spatial impulse response function also is called a point-spread function, i.e., a function that describes the spread of excitation and inhibition around a point input.

Line-spread function. The curves in Figs. 8d, 8e, and 8f describe the line-spread functions of the model; they represent the spread of excitation and inhibition around a thin line stimulus. The intensity $\Delta\ell$ of the line stimulus is such that, when it is centered directly over the receptive field, it provides exactly as much stimulation to the center of the field (w_H) as did the corresponding point stimulus in Figs. 8a, 8b, and 8c. However, line inputs induce more inhibition relative to excitation (by a factor of σ_I/σ_H) than do point inputs.

In the model, only the η ratio of inhibition to excitation matters; it matters not whether it is determined by temporal or spatial factors. Thus, elongating a point stimulus into a line is equivalent to increasing η for a point stimulus by a factor of $\sigma_I/\sigma_H = 3$, thereby greatly enhancing the manifestations of inhibition. For example, the peak of the Phase 2 excitatory response to a point stimulus is six times greater than the peak of the inhibitory response; the peak of the Phase 2 excitatory response to a line stimulus actually is smaller than the peak of the inhibitory response.

Lines are one-dimensional impulses in a two-dimensional space. The line spread functions are the one-dimensional impulse responses, and are the appropriate ones to consider in analyzing one-dimensional stimuli, such as those used to study grating acuity and Mach bands. The acuity target that has been used for the most thorough and analytical experimental studies is the spatial sine-wave grid; responses of the model to these grids are considered later.

Half-field responses. A half-field test stimulus at Location x, is defined as being of retinal illuminance $\Delta\ell$ everywhere to the right of x and of illuminance zero to the left of x. The response of the model to a half-field test is illustrated in Figs. 8g, 8h, and 8i. Because the response has no significant symmetries, it is necessary to calculate the entire response, instead of merely half of it as for points and lines. The intensity $\Delta\ell$ of the half-field stimulus is such that, when the light area completely covers the center of the receptive field, it delivers as much light to w_H as do the corresponding point and line stimuli in the left and center of Fig. 8. At high values of background ℓ, this value of $\Delta\ell$ is such that $\Delta\ell/\ell = 0.25$, a rather low contrast ratio.

Spreading a line stimulus into a half-field greatly enhances the manifestations of inhibition. For example, at high background retinal illuminances, the Phase 2 response to a half-field stimulus is negative everywhere, corresponding to a negative afterimage.

Half-field stimuli are among those that produce the Mach band illusion. Mach bands in the model are evident in Phase 1, Phase 2, and in the steady-state (Figs. 8g, 8h, and 8i), although the vertical scale of the figures is too small to show them well. The transient, Phase 1 Mach bands are less prominent than the steady-state bands. Figure 8h (high background ℓ) illustrates inverted Mach bands, such as would be seen in a blurred afterimage. Mach bands are considered in more detail later.

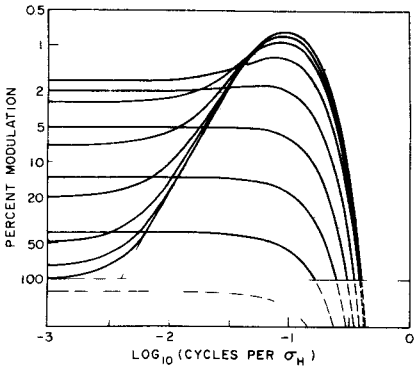

Fig. 9. Spatial sine-wave modulation thresholds of the feedforward section of the model as a function of spatial frequency. Average retinal illuminance is the same for all points of a curve, and differs by 0.5 $\log_{10}$ units for adjacent curves. Dashed curves are theoretical calculations for modulations greater than 100%. Maximum resolution occurs at highest average illuminances for gratings of approximately 0.1 cycles/σ_H.

Responses to Spatial Sine Waves

When the eye is confronted with a sine wave acuity grating (see Section IId), detection of the grid depends on three main variables: the fineness of the grid (spatial frequency ω in terms of number of cycles per degree), the average luminance (ℓ), and the modulation amplitude of the sine perturbation (m, a fraction of ℓ). Detection also depends on other factors, such as the O's internal threshold criterion and the number of cycles viewed. Predictions of the model (considering only the feedforward component) were generated from Eq. 25 by assuming a detection threshold of $\epsilon = 0.01$ and neglecting the effect of number of visible cycles. The predicted sine wave thresholds (m vs ω curves for various values of ℓ) are illustrated in Fig. 9.

Figure 9 shows that as average luminance increases, so does the model's acuity for sine wave gratings. At high luminances, this trend reverses itself for low-frequency gratings, thereby producing a pronounced peak of sensitivity at about 0.09 cycles/σ_H. At the peak, a modulation of about 0.007 is detected. Beyond this peak, the loss of acuity with frequency is very rapid, so that the highest frequency that can be detected is about 0.4 cycles/σ_H.

The modulation transfer functions of Fig. 9 resemble transfer functions measured for the human eye under various conditions, e.g., by Schade (1956), Westheimer (1960), Campbell and Green (1965), Green and Campbell (1965), Robson (1966), Campbell et al (1966), and Campbell and Robson (1968). To determine the amount of quantitative agreement, σ_H must be specified. Perhaps the most easily specified feature of the human detection data is the peak in the transfer function which, at high luminance, occurs at about 8 cycles per degree (Campbell & Green, 1965; Campbell et al, 1966) and at 2 to 4 cycles per degree (Campbell & Robson, 1968). Setting the peak of model's response to 4 cycles per degree determines $\sigma_H = 1.47$ min. The predicted acuity limit is then about 4.4 times the peak (18 cycles per degree) whereas the observed acuity limit is about 50 cycles per degree for stimuli viewed through the normal optics of the eye and about 60 cycles per degree for retinal stimuli of 100% contrast modulation. Typically, the observed acuity limit exceeds the observed sine-sensitivity peak by a factor of about 4 to 20.

Generally, then, the model predicts a too-small range of peak-to-cutoff in sine-wave modulation thresholds. This is not a critical error because it results from the steep high-frequency attenuation characteristic of the normal distribution assumed for $w_H(r)$. In fact, from the observed high-frequency attenuation characteristic, it is possible to calculate a $w_H(r)$ (different from a normal distribution) that would reproduce the observed high-frequency attenuation characteristic exactly and thereby eliminate the prediction error completely.

Multiple sizes of receptive fields. The calculation of an optimally fitting $w_H(r)$ is not made here because the probable reason for the failure of the model to reproduce the sine-wave threshold data precisely is

not its nonoptimal choice of $w_H(r)$ and $w_I(r)$ but its attempt to fit all the data with but a single choice of w_H and w_I. There are now at least three lines of evidence to indicate that various sizes of receptive fields operate jointly to determine thresholds.

Originally, the existence of two or more parallel detection systems with different size constants was proposed by several authors, based on their analyses of detection thresholds (e.g., Hallett, 1963; Campbell & Robson, 1968). Unfortunately, various complex technical factors, such as lack of perfect radial symmetry of grid detectors (e.g., Campbell et al, 1966), the effect of the number of visible grid cycles (Coltman & Anderson, 1960; Campbell & Robson, 1968; Nachmias, 1968) and the fact that low-frequency sine waves extend over nonhomogeneous retinal regions, make it difficult to prove or disprove the multiple receptor hypothesis by simple sine-grid detection experiments.

The strongest evidence for multiple receptor systems comes from Blakemore & Campbell's (1968) study of selective masking of sine-grids by sine-grids. Blakemore observed that prolonged viewing of a particular frequency of sine-wave grating selectively increases the detection threshold for a grating of exactly the same spatial frequency. In Blakemore's analysis, there are at least as many sizes of neurons as there are frequencies that can be selectively masked, and this appears to be a very large number indeed.

Indirect evidence for multiple detection systems. A third kind of evidence for the existence of at least two parallel detection systems, with different parameters, is derived from studies that show that critical duration depends on the O's task; i.e., on whether he detects acuity targets or brightness (Kahneman & Norman, 1964). Additional evidence comes from some unpublished experiments by the author. Critical duration was measured under conditions in which the boundaries of the test field coincided with the boundaries of the masking field, and also in conditions in which the test's boundaries lay within those of a larger masking field. When the T and M boundaries coincide, detection is restricted to detection of a temporal illuminance pulse. Under these conditions, critical duration was found to be substantially smaller than when test boundaries were available for contrast detection. The results imply that a system with a long time-constant is operative when a boundary is being detected and a system with a short time-constant is operative when temporal pulses are being detected.

In terms of information processing, it makes good sense that a system for detection of spatial detail should gather information from a small area (in order to gain maximum spatial resolution), and for a long duration (in order to gain maximum statistical significance of its input). On the other hand, a system for detecting temporal variation must sample information in relatively short periods of time (in order to achieve temporal resolution) and, therefore, in order to obtain a statistically significant sample, it would have to increase the size of the sampled area.

In most normal viewing situations, both the primarily temporal and the primarily spatial systems are likely to be stimulated. In certain conditions, however, even when boundaries are available for spatial contrast detection, they may not be used; that is, detection may be only of the temporal variation. This fact accounts for some obvious discrepancies between apparently similar studies of detection (Sperling & Sondhi, 1968, p. 1143).

Second-Order Threshold Effects

Change of area of integration with light adaptation. The spatial responses of the feedforward system to brief- and to long-duration pulses of light are indicated by the Phase 1 and steady-state responses in Fig. 8. To find the output G(x,y) for small-increment test signals $\ell_T(x,y)$ other than points, $\ell_T(x,y)$ is convoluted (i.e., weighted) by the point-response G from Fig. 8. This procedure uses the property of small-signal linearity in the model. Alternatively, G(x,y) for the test may be calculated directly. When the difference between max G(x,y) and min G(x,y) exceeds the detection threshold ϵ, detection is signaled.

To compare the model with the results of threshold experiments, it is more convenient to reverse the procedure outlined above, and to calculate the spatial impulse response $\hat{G}(x,y)$ implied by the psychophysical threshold results, and then to compare $\hat{G}(x,y)$ with G(x,y). Calculations of this kind (Kincaid et al, 1960; Blackwell, 1963)[8] indicate that the functional impulse response G becomes narrower with light adaptation, much as does G in Fig. 8, i.e., qualitative agreement with the model. The predicted shrinkage of the critical area of the spatial response is given by Eq. 19c. Shrinkage depends on the area factor a. As the area factor a is altered by η to $a' = a\eta$, shrinkage also depends on the effective η value of the test flash used to measure it. For $\eta = 1$, the predicted shrinkage with light adaptation was a factor of ½, which is quantitatively somewhat less than typically was observed by Blackwell (1963).

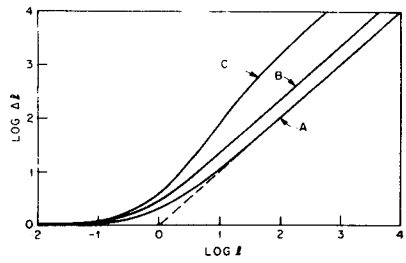

Fig. 10. Dependence of predicted test threshold $\Delta\ell$ on the area of test stimulus. Curve A is for a point stimulus; Curve B is for a half-field stimulus. For A and B, detection is assumed to occur in steady state. Curve C represents theoretical thresholds of a half-field stimulus when detection is assumed to occur in Phase 2. For ease of comparison, the curves were moved vertically so as to make the thresholds superimpose at left (i.e., in the dark).

Pulse duration. Barlow (1958) demonstrated that pulses of a long duration imply a narrower spread function than do brief pulses. This effect is predicted by the model because η for long pulses is larger than η for brief pulses, thus more shrinkage (cf. Eq. 19c, steady-state vs Phase 1 responses in Fig. 8).

Pulse area. A third experimental observation is that large-area stimuli have shorter periods of temporal integration than do small-area stimuli. This effect is predicted by the temporal version of this model (Sperling & Sondhi, 1968) via the assumption that the feedforward controlling signal comes from a larger area (the surround area of the receptive field) than does the controlled signal (center area of the receptive field).

A fourth observation (Barlow, 1957) is that thresholds of large-area stimuli increase from their dark-adapted value at lower luminances than do thresholds of small-area stimuli. This is predicted from Eqs. 12 and 18. The model's prediction is illustrated in Fig. 10, which shows $\Delta\ell$ vs ℓ threshold functions for a point test stimulus and for a half-field test stimulus. The parameters for the curves are $\eta = 1$, $\epsilon = 0.1$.

Figure 10 also illustrates a speculative $\Delta\ell$ vs ℓ function of a large test field for which it is assumed $\eta = 2$, i.e., detection occurs in Phase 2 of a long test flash. This function is particularly suggestive because in a middle range of illuminances, $\Delta\ell$ increases more rapidly than ℓ, a mystifying but not uncommon phenomenon of experimentally obtained $\Delta\ell$ vs ℓ curves. The phenomenon occurs in the model because of the combined changes in sensitivity and in spatial field

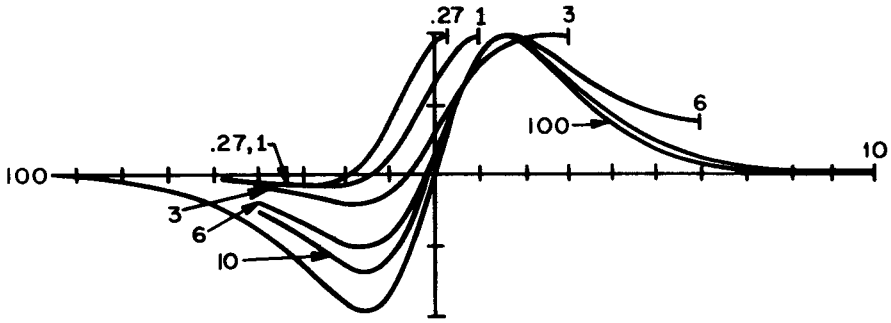

Fig. 11. Spatial response of the model to incremental disks of various diameters. Background retinal illuminance is very high. The responses are superimposed so that the left edge of all the disks coincides with the center vertical line. Radii of the disks are indicated in σ_H units. The maxima of the responses have been normalized to be equal. The horizontal scales are marked in units of one σ_H. Responses are drawn only up to the center of the disk (except for r = 100 σ_H); each response is mirror-symmetrical to the right of its center.

characteristics. The curve for $\eta = 2$ is speculative because when $\eta = 2$, the model predicts that in the interior of an incremental test the response ΔG is less than it would have been in the absence of the test. This condition describes a negative afterimage, and frequently negative afterimages of a test stimulus are more detectable than the stimulus (Brindley, 1959; Sperling, 1960, 1965, p. 556). However, the conditions under which $\Delta\ell$ increases more rapidly than ℓ probably are not limited to detection of afterimages.

The principle underlying the adaptation, area, and duration effects is that the more inhibition (feedforward control signal) there is in the system, the sharper is the spatial spread function and the shorter the temporal function. In the model, these properties result because the inhibitory spatial weighting function w_I is wider than w_H, and because inhibitory control is slower than the controlled signal (Sperling & Sondhi, 1968). Thus, large areas and long-duration stimuli favor inhibition and produce the results characteristic of more inhibition. In fact, the various complex second-order interactions of duration, area, and background luminance are subsumed in a single term [$c_3(\ell)$, Eq. 12], which can be interpreted as the amount of inhibition in the feedforward control pathway at the instant of detection.

Thresholds of Disk-Shaped Test Stimuli

Although disk-shaped stimuli are the most common psychophysical test stimuli, their thresholds are extremely complicated to calculate in the model. Threshold predictions were obtained under the following assumptions: (1) the background retinal illuminance is large; (2) the threshold is small ($\epsilon \ll 1$); and (3) w_H, w_I are normal distributions with $\sigma_I/\sigma_H = 3.0$. The results are illustrated in Fig. 11, which shows the model's output G vs the radial distance from the edge of the disk, for several sizes of disk. Note how the shape of G depends on disk size; in particular, how appreciable minima in G occur only for quite large disks, and how the separation between G_{min} and G_{max} increases as disk size increases.

From the locations and values of G_{min} and G_{max} it is possible to calculate test thresholds $\Delta\ell$. The results of these calculations are illustrated in Fig. 12, which shows how threshold depends on the area ($A = \pi r^2$) of a disk stimulus. Figure 12 shows thresholds calculated at very high background retinal illuminances and also at zero background retinal illuminance; the two sets of thresholds are scaled to coincide for point tests stimuli.

As in psychophysical experiments, for small areas of disk, the threshold depends only on the total energy ($\Delta\ell \propto r^{-2}$, left-hand asymptote). For large disks, $\Delta\ell$ is constant, independent of r. The intersections of the two asymptotes determine the critical areas A_c of the disk stimuli. In the dark $A_c = W_H$ (the effective area of the excitatory center), corresponding to a disk radius of $r = \sigma_H\sqrt{2}$. In the light adapted model, $A_c = .47\ W_H$. These graphically extrapolated values are the same as those obtained by direct calculation from Eq. 19c.

A shallow minimum in $\Delta\ell$ vs r occurs when the radius of the test disk is 2.3 σ_H. The minimum is about 0.2 $\log_{10}$ units deep. The disks which contain the minimum threshold energy, however, have infinitesimally small areas ($r \approx 0$).

There are three basic reasons to expect, in psychophysical data, the depth of observed relative minimum of $\Delta\ell$ vs r to be even less than the depth of the minimum predicted in Fig. 12: (1) multiple sizes of receptive fields, (2) failure to use high background light levels, (3) statistical fluctuation effects.

(1) *Multiple sizes of receptive fields.* The evidence for multiple sizes of receptive fields was presented earlier in connection with sine-wave thresholds. The effect of altering the scale of w_H and w_I would be to shift the curves of Fig. 12 laterally. (Altering the sensitivity of a model would correspond to shifting the curves of Fig. 12 vertically.) When disk-thresholds are determined by the joint action of receptive fields of various sizes and sensitivities, the predicted overall $\Delta\ell$-vs-radius function is obtained by averaging the sensitivities of the (shifted) individual functions. The overall function, perforce, has a shallower minimum than any of the individual functions. In fact, with certain distributions of component receptive field sizes, the overall $\Delta\ell$-vs-radius function will have no observable minimum. As a practical matter, it would be impossible to discriminate psychophysically between a shallow minimum and no minimum at all. Differences occur only with large-area disks; the gross nonhomogeneity of the retina makes such threshold data too difficult to interpret.

(2) *Background retinal illuminance.* If ℓ is not asymptotically high, individual $\Delta\ell$-vs-radius functions will be intermediate between the extremes illustrated in Fig. 12; the relative minimum will be proportionately shallower.

(3) *Statistical fluctuation effects.* There are various statistical factors that enter into psychophysical thresholds. For example, detection may occur independently in different parts of the test field. The number of opportunities for detection

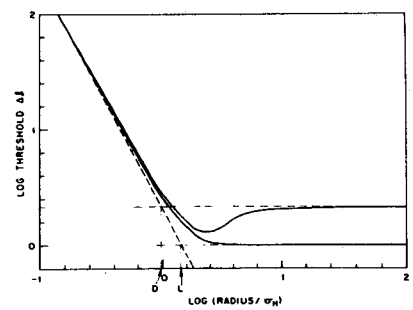

Fig. 12. Predicted threshold $\Delta\ell$ of an incremental disk stimulus as a function of its radius. Upper curve is for high background retinal illuminances; lower curve is for the dark-adapted model. For ease of comparison, the curves have been moved vertically to superimpose at the extreme left, i.e., for point test stimuli. The radii of disks of critical area are indicated by (D) dark-adapted and (L) light-adapted.

would be proportional to the circumference of the test disk (i.e., to $2\pi r$) for $r > \text{diam}(w_H)$. Another factor is the quantum fluctuation of disk energy, even when $\Delta\ell$ nominally is constant. Quantum fluctuations increase in proportion to the square root of the energy; the relative fluctuation therefore decreases with the square root of energy. In a detection experiment, the decrease in relative fluctuation could yield an improvement in detection proportional to $A^{1/2}$. The slope at the right-hand side of the minimum of $\Delta\ell$ (Fig. 12) is less than that of $A^{1/2}$ (slope = 1) so that if either of these statistical effects were operative in the region of the minimum, the minimum would be obscured entirely. In fact, in psychophysical experiments, threshold $\Delta\ell$ usually continues to decrease with disk diameter, even for quite large disks, presumably for the reasons given above.

Basically, the reason that the inhibitory field does not produce a large $\Delta\ell$-minimum in the predicted disk-detection function is that large disks are detected not in their center but at their boundaries. The presence of the $\Delta\ell$-minimum has been a traditional psychophysical criterion of physiological inhibition. The small, shallow minimum predicted by the model, even at the highest background retinal illuminance, plus the additional facts that the multiplicity of receptive field sizes and statistical effects would tend to obscure it, account for the difficulty of observing evidence for inhibition in psychophysical detection experiments.

When critical areas are calculated from the thresholds of disks viewed foveally in absolute darkness, they typically have diameters of 4 to 5 min (Bouman & Walraven, 1957; Glezer, 1965).[9] To fit the model to these data, σ_H is estimated from $\sigma_H = \text{diam (critical area in dark)}/\sqrt{8}$, giving a σ_H of about 1.4 to 1.7 min. The value of σ_H estimated from disk detection experiments thus is in agreement with the value of 1.47 estimated from the peak of the observed sine-wave threshold function.

Mach Bands

Figure 8 showed that the output of the model has a much shallower minimum for a point stimulus than it does for a line stimulus, and the minimum for a line is shallower than for a half-field. If the height of the uninhibited maximum is designated as h, then the depth of the minimum is approximately proportional to h/σ^2 for a point, to h/σ for a line and to h directly for a half-field. Since the minima of G are the signs of inhibition, to demonstrate inhibition one should use stimuli containing relatively large uniform—or nearly uniform—areas. The stimuli for producing Mach bands are precisely such stimuli. Furthermore, the value of the illuminance difference between adjacent areas of the stimulus determines ΔG in the output of the model. Thus, by using test stimuli that are well above threshold, inhibitory effects are magnified in the output. With such stimuli, the model predicts quite prominent Mach bands; these were illustrated in Fig. 2.

Other predictions about Mach bands by the model also are qualitatively in agreement with observations. The model predicts that Mach bands are not observed at extremely low retinal illuminances, that at moderately low illuminances the bright Mach band is more prominent than the dark band, while at high illuminances the reverse is true. It also correctly predicts that Mach bands can be seen in brief flashes, but that they are less prominent than bands that appear in continuous viewing.

The main deficiency of the model is in its predictions of apparent brightness of large uniform areas. Once a uniform area reaches a moderate intensity, it no longer can induce significantly greater responses in the model as its intensity is increased further; i.e., the model's response saturates quickly. Thus the model can predict brightness phenomena that involve low and moderate degrees of contrast but not brightness phenomena in general. In the case of Mach bands, this means that predictions of the model tend to emphasize the prominance of a band relative to the brightness of adjacent uniform areas. It is quite probable that, in the visual system, information about overall brightness is handled differently than information about contours, i.e., about brightness differences. Insofar as this is true, the deficiency of a contrast-detection model in predicting overall brightness of uniform areas is not a defect.

The Effect of Feedback in the Model

Since spatial effects (as opposed to adaptational effects) at the level of the feedback stage (FB) are very difficult to distinguish from those at the level of the feedforward stage (FF), in the foregoing analyses, all spatial effects were assumed to occur in the FF stage. We now consider the conditions under which FB becomes important.

In predicting disk-thresholds (Figs. 11 and 12), FF alone and FB + FF are equivalent because the thresholds are determined at a constant adaptation level ℓ and for infinitesimally small increments $\Delta\ell$.

In predicting sine-wave contrast thresholds and Mach bands, FF alone and FB + FF are similar—but not exactly the same—because adaptation varies somewhat across the field.

When predicting the effect of background retinal illuminance on threshold values of $\Delta\ell$, FF and FB + FF are quite different; here the difference was explicitly considered (Fig. 6).

The most important effect of FB is in its temporal control of adaptation. This was subsumed in the parameter η. We consider now a significant phenomenon of contrast detection that, strangely, appears to have been overlooked. In very brief flashes, Os cannot detect contrast differences of less than about 5% across a boundary; 10% is a more common limit on performance (Brindley, 1959; Sperling, 1965). In a long flash of 1.0 sec duration, Os easily detect a contrast of 2%. Os will fail to detect a low-contrast boundary in a brief flash and yet detect the same boundary in a long flash that contains but a fraction of the energy. Better detection of long flashes (when energy is held constant) is the opposite of temporal integration. In terms of the model, it implies that the output "saturates" when the input becomes large. This kind of saturation, of course, is a characteristic of the FB stage. However, to predict this phenomenon quantitatively, indeed, to predict any of the vast number of complex temporal-spatial interactions in vision, would require a much more complete dynamic specification of the FB and FF stages than was necessary for the elementary model of spatial vision proposed here.

IV. SUMMARY AND CONCLUSIONS

A three-component model for spatial vision is proposed. The first component is an RC stage controlled by shunting feedback. The second component is an RC stage controlled by shunting feedforward, with the controlling feedforward signal originating from a larger area than the controlled signal. Finally, a detector compares the locations that produce the largest and the smallest outputs for a particular signal and indicates detection when the algebraic difference exceeds a threshold criterion ϵ. The shunting stages are analogues of synaptic inhibitory processes.

The model predicts that thresholds for increments $\Delta\ell$ on a background ℓ follow a square-root law at low background luminance ($\Delta\ell \propto \ell^{1/2}$) and a Weber law at high background luminance ($\Delta\ell \propto \ell$). The model achieves its Weber-law response as a consequence of the feedforward component and without having any signals proportional to the logarithm of the stimulus. In the model, square-root law responses are a consequence of the feedback component, not of quantum

noise directly; the square-root law is presumed to be an evolutionary adaptation (FB-component) to quantum noise.

With two stimuli A, B, both mask a test $\Delta\ell$, the model predicts that the combined threshold equals the larger of $\Delta\ell_A$ and $\Delta\ell_B$ at low luminances and equals $\Delta\ell_A + \Delta\ell_B$ at high luminances; i.e., masking effects add linearly in the model at high luminances.

The model predicts a receptive field structure that, in the dark, contains only an excitatory center and that, in the light, contains an inhibitory surround, the apparent center reducing in size with increasing background illumination.

By using a single parameter to characterize the temporal response of the model [derived from a similar system by Sperling & Sondhi (1968)], the model predicts different receptive fields for short- and for long-duration pulses. It correctly predicts the following second-order threshold effects: (a) reduced spatial integration for long-duration pulses, (b) reduced temporal integration for large-area pulses, (c) greater effect of background retinal illuminance on detection of large-area pulses.

When disk test stimuli are superimposed on high background retinal illuminances, the response of the model to the disk varies greatly with disk diameter. For example, the model detects disks of diameter smaller than the diameter of its inhibitory surround at their center, and detects larger disks at their boundaries. This explains why disk-detection experiments usually fail to indicate an inhibitory surround. The model accounts for Mach bands and shows why they are an ideal way to demonstrate spatial inhibition.

The model gives only fair predictions of sine-wave contrast thresholds. By modifying the normally distributed spread function that was assumed for excitation, the model could virtually eliminate its prediction errors. However, the source of difficulty is shown to be the existence of units of different sizes and of different temporal characteristics; the same basic model (but with a different spatial scale) applies to each size of unit.

REFERENCES

BARLOW, H. B. Increment thresholds at low intensities considered as signal/noise discriminations. Journal of Physiology (London), 1957, 136, 469-488.

BARLOW, H. B. Temporal and spatial summation in human vision at different background intensities. Journal of Physiology (London), 1958, 141, 337-350.

BARLOW, H. B. A method of determining the overall quantum efficiency of visual discriminations. Journal of Physiology (London), 1962a, 160, 155-168.

BARLOW, H. B. Measurements of the quantum efficiency of discrimination in human scotopic vision. Journal of Physiology (London), 1962b, 160, 169-188.

BLACKWELL, H. R. Neural theories of simple visual discriminations. Journal of the Optical Society of America, 1963, 53, 129-160.

BLAKEMORE, C., & CAMPBELL, F. W. Adaptation to spatial stimuli. Journal of Physiology (London), 1968, 200, 11-13.

BOUMAN, M. A., & WALRAVEN, P. L. Some color naming experiments for red and green monochromatic lights. Journal of the Optical Society of America, 1957, 47, 834-839.

BRINDLEY, G. S. The discrimination of after-images. Journal of Physiology (London), 1959, 147, 194-203.

BRINDLEY, G. S. *Physiology of the retina and the visual pathway.* London: Edward Arnold, 1960.

CAMPBELL, F. W., & GREEN, D. G. Optical and retinal factors affecting visual resolution. Journal of Physiology (London), 1965, 181, 576-593.

CAMPBELL, F. W., KULIKOWSKI, J. J., & LEVINSON, J. The effect of orientation on the visual resolution of gratings. Journal of Physiology (London), 1966, 187, 427-436.

CAMPBELL, F. W., & ROBSON, J. G. Application of Fourier analysis to the visibility of gratings. Journal of Physiology (London), 1968, 197, 551-566.

COLTMAN, J. W., & ANDERSON, A. E. Noise limitations to resolving power in electronic imaging. Proceedings of the IRE, 1960, 48, 858-865.

COOMBS, J. S., ECCLES, J. C., & FATT, P. The specific ionic conductances and the ionic movements across the motoneuronal membrane that produce the inhibitory postsynaptic potential. Journal of Physiology (London), 1955, 130, 326-373.

DeVRIES, H. The quantum character of light and its bearing upon threshold of vision, the differential sensitivity and visual acuity of the eye. Physica, 1943, 10, 553-564.

FATT, P., & KATZ, B. The effect of inhibitory nerve impulses on a crustacean muscle fibre. Journal of Physiology (London), 1953, 121, 374-389.

FUORTES, M. G. F., & HODGKIN, A. L. Changes in time scale and sensitivity in the ommatidia of Limulus. Journal of Physiology (London), 1964, 172, 239-263.

FURMAN, G. G. Comparison of models for subtractive and shunting lateral-inhibition in receptor-neuron fields. Kybernetik, 1965, 2, 257-274.

GLEZER, V. D. The receptive fields of the retina. Vision Research, 1965, 5, 497-525.

GREEN, D. G., & CAMPBELL, F. W. Effect of focus on the visual response to a sinusoidally modulated spatial stimulus. Journal of the Optical Society of America, 1965, 55, 1154-1157.

HALLETT, P. E. Spatial summation. Vision Research, 1963, 3, 9-24.

HARTLINE, H. K., & RATLIFF, F. Spatial summation of inhibitory influences in the eye of *Limulus*, and the mutual interaction of receptor units. Journal of General Physiology, 1958, 41, 1049-1066.

HECHT, S., SCHLAER, S., & PIRENNE, M. H. Energy, quanta, and vision. Journal of General Physiology, 1942, 25, 819-840.

KAHNEMAN, D., & NORMAN, J. The time-intensity relation in visual perception as a function of the observer's task. Journal of Experimental Psychology, 1964, 68, 215-20.

KINCAID, W. M., BLACKWELL, H. R., & KRISTOFFERSON, A. B. Neural formulation of the effects of target size and shape upon visual detection. Journal of the Optical Society of America, 1960, 50, 143-148.

LAMAR, E. S., HECHT, S., SCHLAER, S., & HENDLEY, C. D. Size, shape and contrast in detection of targets by daylight vision. I. Data and analytical description. Journal of the Optical Society of America, 1947, 37, 531-545.

LOWRY, E. M., & DePALMA, J. J. Sine-wave response of the visual system. I. The Mach phenomenon. Journal of the Optical Society of America, 1961, 51, 740-746.

MACH, E. Über die physiologische wirkung räumlich vertheilter Lichtreize, IV. Sitzungsberichte der mathematisch-naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften, 1868, 57/2, 11-19.

NACHMIAS, J. Visual resolution of two-bar patterns and square-wave grating. Journal of the Optical Society of America, 1968, 58, 9-13.

PIRENNE, M. H. Absolute thresholds and quantum effects *and* Quantum fluctuations at the absolute threshold. In H. Davson (Ed.), *The eye.* Vol. 2. New York: Academic Press, 1962. Pp. 123-140; 141-158.

PURPLE, R. L., & DODGE, F. A. Interaction of excitation and inhibition in the eccentric cell in the eye of *Limulus.* Cold Springs Harbor Symposia on Quantitative Biology, 1965, 30, 529-537.

RATLIFF, F. *Mach bands. Quantitative studies on neural networks in the retina.* San Francisco: Holden-Day, 1965.

ROBSON, J. G. Spatial and temporal contrast-sensitivity functions of the visual system. Journal of the Optical Society of America, 1966, 56, 1141-1142.

RODIECK, R. W. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vision Research, 1965, 5, 583-601.

RUSHTON, W. A. H. Increment threshold and dark adaptation. Journal of the Optical Society of America, 1963, 53, 104-109.

SCHADE, O. H. Optical and photo-electric analogy of the eye. Journal of the Optical Society of America, 1956, 46, 721-739.

SPERLING, G. Negative afterimage without prior positive image. Science, 1960, 131, 1613-1614.

SPERLING, G. Temporal and spatial masking. I. Masking by impulse flashes. Journal of the Optical Society of America, 1965, 55, 541-559.

SPERLING, G., & SONDHI, M. M. Model for visual luminance discrimination and flicker detection. Journal of the Optical Society of America, 1968, 58, 1133-1145.

von BÉKÉSY, G. Neural inhibitory units of the eye and skin. Quantitative description of contrast phenomena. Journal of the Optical Society of America, 1960, 50, 1060-1070.

WESTHEIMER, G. Modulation thresholds for sinusoidal light distributions on the retina. Journal of Physiology (London), 1960, 152, 67-74.

NOTES

1. Address: Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07974. (Currently a fellow of the John Simon Guggenheim Memorial Foundation, University College London.)

2. The author wishes to acknowledge the intellectual contribution of G. Furman, who first proposed the possibility of using shunting feedforward networks to explain visual contrast detection.

3. When logarithmic transformations of input intensity are assumed to occur before substractive lateral interaction, subtractive interactions of the logarithms mimic multiplicative interactions of the inputs. The net

effects can be quite similar to the effects of shunting interactions. The logarithmic subtractive approach is unattractive because it requires two complex operations (logarithms and subtraction) to do what shunting processes do naturally in a single interaction. Even in *Limulus*, whose eye represents the most successful case of the subtractive theory of interaction, the subtractive theory is yielding to a shunting analysis (Purple & Dodge, 1965).

4. Mach computed a lateral interaction term of the form H/I, which is nearly equivalent to $H/(1 + kI)$ of Eq. 3a. However, Mach assumed that the output was not given directly by the interaction term, but was proportional to input $\times$ interaction; namely, H^2/I (Mach, 1868, p. 15).

5. The minimum of G(x) in Fig. 8c indicates Δx.

6. See Fig. 8i; also, the zero of G(x) in Fig. 8f gives approximately $\pm\Delta x/2$.

7. See Sperling & Sondhi (1968) for a specific theory of slow inhibition.

8. The convolution procedure used by Kincaid et al (1960) and by Blackwell (1963) assumes implicitly that disk test stimuli are detected at their center. The procedure is invalid when test fields are detected at boundaries. (See next section.)

9. These values are among the smallest critical areas that have been reported. The small values are used here because most experimental artifacts, such as accommodation error (from fixation lights of different wavelengths than the disks being detected) and inaccurate fixation tend to increase observed critical areas.

(Accepted for publication November 20, 1969.)