CHAPTER 7

The beginnings of visual perception: the retinal image and its initial encoding

Appendix: Fourier transforms and shift-invariant linear operators

JOHN I. YELLOTT, JR. | *University of California, Irvine, California*
BRIAN A. WANDELL | *Stanford University, Stanford, California*
TOM N. CORNSWEET | *University of California, Irvine, California*

CHAPTER CONTENTS

SCOPE AND ORGANIZATION

The organizing theme of this chapter is the general question: What aspects of human visual perception can be thoroughly explained in physiological terms at the present time? This introductory section surveys the overall status of physiological explanations of visual phenomena. At present, tight explanatory links between perception and physiology are still largely confined to the level of the optics of the retinal image and its initial neural registration by the photoreceptors. However, understanding of the neural events that immediately follow quantum absorption is growing very rapidly: A number of perceptual phenomena seem tantalizingly close to complete explanation. The subsequent sections provide fairly detailed discussions of three special topics: *1*) high spatial frequency sensitivity or visual acuity (p. 260), *2*) color vision (p. 280), and *3*) light and dark adaptation (p. 293). The first two topics furnish outstanding examples of important visual phenomena that now seem to be thoroughly explainable in physical terms. The third topic is an example of a perceptual phenomenon that seems within reach of current physiological techniques and explanatory principles, but still some distance from a thoroughly satisfactory physical explanation. Finally, a mathematical *Appendix*, p. 302, outlines the Fourier analytic ideas underlying much current work in vision, and in particular the discussion of retinal imagery and spatial contrast sensitivity in our section VISUAL ACUITY p. 260.

PHYSIOLOGICAL EXPLANATION IN VISUAL SCIENCE: THREE CLASSES OF PERCEPTUAL PHENOMENA

The phenomena of human visual perception can be usefully divided into three classes. The smallest and most exclusive class consists of phenomena whose properties can presently be explained in terms of accepted physical principles together with well-established anatomical and physiological properties of the

visual apparatus. In this class one can place two of the major phenomena of color vision—trichromacy at high-illumination levels and monochromacy (total color blindness) at low levels; and also the main features of visual detection of fine spatial detail—technically, the high-frequency end of the spatial contrast sensitivity function (the spatial "modulation transfer function" of the visual system). In addition one could include here various pathological effects having to do with visual field losses produced by lesions in the lower visual pathways (where the anatomical wiring diagrams are uncontroversial) and a few perceptual curiosities such as the blind spot (the 6° gap in the visual field produced by the hole in the receptor mosaic through which ganglion cell axons leave the retina) that can be attributed to straightforward anatomical considerations. The reader may be able to think of a few other candidates, but at the moment we do not think there are very many.

All of these class 1 phenomena can be conceptualized as a loss of information potentially available in the visual stimulus. In the case of color vision the visual system loses wavelength information, so that spectrally different lights are perceptually indistinguishable; in spatial vision the system loses high-frequency modulation information, so that stimuli which differ only in their fine detail cannot be discriminated; and the other cases mentioned above can be characterized in a similar fashion. This preoccupation with information loss is not accidental: The only perceptual phenomena universally acknowledged as being susceptible to explanation in physical terms are those that can be described as sensory losses of information—or equivalently (since information is ultimately carried by differences) in terms of a perceptual inability to discriminate between physically different stimuli. Brindley's (29) statement of the epistomological issue here has been especially influential:

> The main function of science, in those of its branches that have advanced beyond the primitively exploratory stage, is the formulation and testing of hypotheses which have exact and potentially observable implications. The inescapable implications of any hypothesis can necessarily be expressed in terms which appear either in the statement of the hypothesis, or in the background of generally accepted theory assumed in conjunction with it. For physiology, the terms used in stating the theoretical background are physico-chemical and anatomical; so it would seem that no physiological hypothesis that is also stated in physical, chemical and anatomical terms can ever predict the result of a sensory experiment, in which a report of sensations is concerned. There is, however, one class of predictions that can be made if we add to our theoretical background a single hypothesis that is very difficult to doubt. The additional hypothesis required is that whenever two stimuli cause physically indistinguishable signals to be sent from the sense organs to the brain, the sensations produced by these stimuli, as reported by the subject in words, symbols or actions, must also be indistinguishable. (p. 132–133).

Brindley's (29) assessment of other potential explanatory principles is also worth quoting, because it represents the most conservative position, and probably reflects the views of the majority of sensory physiologists:

> If a physiological hypothesis, i.e. a hypothesis about function that is stated in physical, chemical and anatomical terms, is to imply a given result for a sensory experiment, the background of theory assumed in conjunction with it must be enlarged to include hypotheses containing psychological terms as well as physico-chemical and anatomical. These may be called *psycho-physical linking hypotheses.* The one that has already been stated above, namely that physically indistinguishable signals sent from sense organs to the brain cause indistinguishable sensations, is the most general, and at the same time the most difficult to doubt, that has yet been proposed. It seems to me that it is the only one that is at present sufficiently secure to deserve inclusion in the body of generally accepted theory. (p. 134).

This degree of skepticism as to the value of other linking hypotheses is not universally shared, but at present there is apparently no general acceptance of any explanatory principle except the one endorsed here by Brindley.

The distinguishing property of the perceptual phenomena in our first class is that they can not only be characterized in terms of a loss of information, but in addition—and most important—we can confidently pinpoint the anatomical and physiological factors responsible for the specific form of the loss. Thus in the case of color vision, wavelength information is lost at the point of light absorption in the photoreceptors because any given receptor only records the fact that a photon has been absorbed, and not the wavelength of that photon [Rushton (112) calls this the "univariance" principle]. Consequently wavelength discrimination is only possible by virtue of a comparison between the quantum catches of receptors that have different absorption spectra, and the specific quantitative form of the chromatic information loss in human vision can be explained as an inevitable consequence of the limited variety of absorption spectra represented in the human retina: Night vision is monochromatic because all rods share the absorption spectrum of rhodopsin and daylight vision is trichromatic because there are three different cone pigments. In the case of spatial contrast sensitivity, high-frequency spatial information is lost because of the optical limitations of the eye: We cannot see ultrafine spatial details because they are washed out in the physical process of image formation and consequently never appear in the retinal image (though as we shall see in VISUAL ACUITY, p. 260, however, this story is complicated by a subsequent neurally imposed frequency cutoff that nearly—and quite mysteriously—matches the limits imposed by the optics). The anatomical and physiological factors responsible for the losses characterizing our other class 1 phenomena have already

been noted in connection with their original descriptions.

In a second class, one can place a large number of visual phenomena for which one can envision plausible physiological models, but for which the specific anatomical or physiological mechanisms remain uncertain. Most of these can be characterized as information losses and fall into the explanatory framework common to our class 1 phenomena. The outstanding examples of class 2 phenomena are light and dark adaptation, i.e., changes in visual sensitivity due to changes in the mean illumination level. These effects might in principle be entirely accounted for in terms of photoreceptor properties (that is, one could readily imagine a visual system constructed in that fashion), and both psychophysical evidence and single-unit recordings indicate that receptors play a very substantial role in light adaptation: Changes in their sensitivity certainly account for a significant share in the overall perceptual effect. In the particular case of the phenomenon of rod saturation (the inability of rod-mediated vision to detect contrast at high-illumination levels) it has been suggested that the perceptual effect may be entirely explained in terms of a limiting of the rod outer segment membrane potential due to the closure of all of its sodium-permeable channels (97). However, in general it is clear that more proximal cells adapt to the prevailing visual environment and require some time to readjust when that environment changes. (12, 46) At the present time not enough is known about the quantitative properties of retinal circuits to support any definitive allocation of responsibility for the overall adaptation effect; this is an important target of current investigation.

The same can be said of many temporal phenomena; notably the inability to detect very rapid temporal modulation (flicker fusion), which might in principle by entirely explained in terms of the temporal smoothing properties of receptors—again, in the sense that one could construct plausible models on this basis, and also in the sense that receptors are known to act like temporal filters (26) and certainly play an important role in damping out rapid flicker. However, there is substantial evidence that "photopic flicker sensitivity is normally controlled, not by cone cells, but by the interactive pathways of the retina" (77).

Many other loss-like phenomena could also be placed in this second class, e.g., the subjective disappearance of stabilized retinal images, the suppression of low-frequency spatial contrast (together with such associated phenomena as mach bands), visual masking effects (28), various pattern and motion aftereffects [e.g., grating adaptation effects of the sort reported by Blakemore and Campbell (21), and Purkinje's waterfall illusion], and some aspects of depth perception [those related to pathological deficits in binocular stereopsis caused by developmental abnormalities that eliminate the disparity sensitive cortical neurons described by Barlow et al. (11)]. In all of those cases the major barrier to full understanding is not the inability to construct plausible physiological models that could in principle account for the phenomena, but rather the lack of a sufficiently rich physiological data base to force consensus on a single specific model. What is needed, of course, is a detailed quantitative understanding of neural circuits, and especially of neural signals as they are actually transmitted, i.e., by transmitter substance release or electrical coupling. Studies aimed at obtaining information of that sort have only recently become technically possible (see, for example, ref. 12), and in fact it has only lately been established that receptors reduced their rate of transmitter release in response to light (102). This central fact is perhaps not surprising, since the membrane response to light is hyperpolarization, but it does pose a new challenge for modelers and also presents the functionally curious spectacle of a receptor-signaling mechanism that works hardest in the dark, when it has nothing to report.

Finally in connection with class 2 it is necessary to say something about heterochromatic brightness, which is the classic example of a visual phenomenon that cannot readily be described in terms of information loss, but instead both requires and suggests a physiological explanation based on some other principle. Intuitively it seems clear that lights can be at least roughly ordered on a dimension of "brightness" even when their colors are different—for example, one has no difficulty in judging that the sun is brighter than the sky—and the precise quantitative relationship between the physical properties of lights and their psychological classification in terms of brightness has been a long-standing psychophysical problem (78). Early experimental work suggested that brightness exactly satisfied an additive relationship: If light A matches light B in brightness, and C matches D, then the combination A + C (i.e., the physical superimposition of lights A and C on the retina) must match B + D [Abney's law (1)]. This additivity hypothesis has important practical implications in photometry and is officially embodied in applications of the Commission Internationale de l'Éclairage (C.I.E.) photopic luminosity curve (the "photopic standard observer") which supplies the spectral weighting function for automatic photometers. The extent to which it actually holds appears to depend on the nature of the brightness judgment required of an observer. When the observer is simply asked to make a direct brightness comparison between isolated steady lights, Abney's law fails quite dramatically (60). When brightness is measured by minimizing apparent flicker against a standard, the deviations from additivity are significant, though apparently tolerable for industrial standards (71). When measured by the relatively new technique of minimizing apparent contrast at a border (25, 127), additivity appears to hold within measurement error (71).

Suppose Abney's law is valid for some psychophysical context: How could one account for it physiologi-

cally? Clearly this is not a straightforward case of sensory loss of information potentially available in a physical stimulus, since the observer is not incapable of discriminating between heterochromatic lights equated for brightness—on the contrary, his task is to overlook their color differences and base his judgment on some perceptual dimension on which they are the same. What is at issue then is not an inability to discriminate, but rather an ability to classify discriminable stimuli in a systematic way. How can such an ability be brought within the scope of physiological explanation?

A first step is to determine the neural coding demands implied by the ability to classify stimuli in this particular fashion—in other words, what kinds of numbers (i.e., neural signals) would a mechanism have to assign to stimuli in order to match the observed psychophysical classification? The precise form of the relational structure implied by Abney's law has been informally understood for some time, and recently has been spelled out rigorously in two papers by Krantz (79, 80) which apply modern ideas of fundamental measurement theory to the classic problems of color vision. From the standpoint of physiological modeling, the essential result is that any device which classifies lights according to Abney's law must base its decision on a linear combination of the trichromatic color-matching coordinates—a condition that would be satisfied most naturally in the retina by a mechanism that based its output on a weighted average of the quantum catch in all three cone systems.

A second step is to imagine a sensory "channel" that carries the signal computed by our hypothetical brightness mechanism: operationally such a channel is defined simply in terms of the assumption that its response to two lights is the same whenever they produce the same value on the brightness dimension—e.g., the same value of weighted average quantum catch. Therefore, the state of this channel can be completely specified by a single real number. Physiologically, such a channel might be identified with the nonopponent ganglion cells found in primate retinas (42). Of course this channel is assumed to exist in parallel with other channels carrying information about other properties of the stimulus (notably, opponent ganglion cells carrying chromatic information), and on those channels the responses to two stimuli of equal brightness will generally not be identical. However, one can suppose that when brightness judgments are called for, the observer can somehow base his decisions solely on the output of the brightness channel, ignoring information on the other channels.

Now if such channels really existed (and an enormous amount of current work on many problems assumes more or less explicitly that they do), we could think of brightness in terms of a sensory loss along one channel, and in this way the psychophysical linking hypothesis described by Brindley would still in a sense apply. However, it will not apply strictly unless one can somehow manage to silence all the other channels, so that the observer's discriminations have to be based on the brightness channel alone. Unless this can be arranged, there remains an uncomfortable nonphysical component in the theory, namely the fact that we cannot predict what an observer *must* do in a brightness judgment task because we do not know what factors determine his ability or willingness to base his judgments on the proper channel. This nonobligatory "psychological" aspect of such multichannel models is epistemologically disturbing. Nevertheless, modeling along such lines seems to present the only obvious line of progress, since not many visual phenomena can be clearly characterized in terms that strictly fit the linking hypothesis quoted earlier. What seems likely is that converging evidence will gradually increase the apparent tangibility of certain specific sensory channels, so that it will eventually come to seem natural to regard them as legitimate entities in a physical theory—notwithstanding the possible psychological foibles of a still physically elusive human "observer" who reads their output. This in fact is already happening in the case of the brightness channel: There is reason to believe that this channel can be isolated by means of rapid flicker, so that discrimination tasks can be arranged that are only possible on the basis of its output, the chromatic channels being irrelevant (77). Conversely, the brightness channel itself can be silenced by requiring a discrimination between stimuli equated for brightness, and this results in rather dramatic perceptual effects (e.g., ref. 58a), which strongly support the idea that some fundamental cleavage in the visual system is being revealed.

Finally, in a third class one can place all those visual phenomena for which we have no clear conception of even the outlines of a physiological explanation. Most prominent here are phenomena involving memory—pattern recognition abilities in particular, but also such purely subjective phenomena as visual imagery and visual experience in dreams and hallucinations. Most aspects of three-dimensional spatial perception (especially monocular depth perception) should also be included here, along with most aspects of motion perception. In all these cases the phenomena cannot be characterized in terms of information loss along well-established visual pathways, and we have no clear understanding of how they might be related to neural mechanisms. Consequently the shape of their ultimate physiological explanations cannot easily be foreseen at the present time.

The following handbooks, textbooks, and review articles may be useful in gaining an overview of the current state of visual science: references 12, 35, 40, 50, 62, 66, 68, 73, 81, 90, 99, 106, 144.

VISUAL ACUITY

Overview

One of the most obvious aspects of visual experience is our limited ability to resolve spatial detail: If a

black-and-white grating is made fine enough, it cannot be distinguished from a uniformly gray field. (At 25-cm viewing distance one can just tell the difference when the width of each bar is 40 μm.) Early in the 17th century the Spanish physician Daza de Valdes (in ref. 84) measured the spatial resolution ability of his patients by having them count grains of mustard in a line; by the end of that century Phillipe de La Hire had "calculated that the limiting fineness of vision was 1/8000 inch on the retina" (which corresponds to a grating in which each stripe subtends 0.6 min of visual angle, i.e., 0.83 cycle/min), and attributed this limit to the structure of the retina: ". . . such is the smallness of the net of which it is made up" [(84), p. 107]. Modern research on visual acuity, influenced by ideas from information theory and Fourier optics, has concentrated on the ability to detect sinusoidal spatial modulation, and has attempted to determine the specific limitations imposed first by the optical apparatus of the eye, and subsequently by neural factors. Under normal viewing conditions it turns out that smearing of the retinal image due to optical factors can account for the observed upper limit of resolution capacity [which according to contemporary measurements is essentially the same as that reported by de La Hire (in ref. 84), i.e., roughly 1 cycle/min]: The optics of the eye simply cannot transmit spatial information at higher frequencies. Consequently one can say that the invisibility of spatial detail beyond this resolution limit is due to the optics of the eye: No matter how the retina and subsequent neural stages of the visual system are designed, they cannot extract spatial information from a visual stimulus unless it is available in the retinal image. In this sense, the upper limits of the spatial resolution capacity of the visual system under normal viewing conditions can be regarded as well understood.

However, this simple story is complicated by the fact that local acuity outside the center of the retina—and also at the center under special conditions that eliminate the normal optical blurring—cannot be accounted for in terms of optical factors, but instead requires a neural explanation. At present not enough is known about the details of neural interactions to allow one to construct definitive models of spatial information processing in the visual system: A great many important facts have been discovered by single-unit recording and other modern techniques, but one still cannot confidently assign any specific aspect of spatial contrast vision to a specific neural mechanism. In this sense, firm understanding has still not progressed beyond the level of the physical properties of the retinal image. In fact it does not appear that we fully understand the general principles underlying the neural processes that limit visual acuity: In *Neural Limits of Visual Acuity*, p. 275, we point out that current physiologically motivated models for spatial contrast detection do not explain how the visual system is able to suppress the counterfeit spatial frequencies that should be generated by its relatively coarse sampling of the retinal image. In a sense, we still do not understand the role played by the smallness of the retinal net.

This section first reviews the optical properties of the eye, with particular emphasis on factors that degrade the quality of the retinal image. Then we discuss the psychophysics of human spatial contrast sensitivity and attempt to relate these results to the optics of the eye and the physical and neural properties of the retina. The discussion centers around concepts derived from information theory and Fourier optics (27, 51, 62, 96); the mathematical details here are summarized in the *Appendix* to this chapter, p. 302. Other recent review articles related to the topic of this section are those by Westheimer (140), Thomas (124), Ripps and Weale (103), and Kelly (74). Le Grand's (84) classic text provides an extensive and authoritative discussion of visual acuity, and a recent theoretical paper by Snyder et al. (118) describes an interesting calculation of the spatial information capacity of the retina that takes into account quantum noise effects as well as the optical and anatomical factors considered here. [We do not deal explicitly with quantum noise because we are concerned with the ability to see fine spatial detail under optimal conditions, which means at illumination levels high enough to guarantee very high quantum signal-to-noise ratios in the retinal image. For general discussions of quantum noise effects in vision, see Barlow (10a) and Rose (108.)]

The Retinal Image

Strictly speaking, the term "retinal image'" is ambiguous, because from an optical standpoint the thickness of the retina is quite appreciable—an image in focus on its vitreal surface can be as much as 2 diopters out of focus at the outer segments of the photoreceptors, where vision is actually initiated. Throughout this discussion we use "retina" to mean the outer segment layer and "retinal image" to mean the image formed at that level of the retina. To understand the limits of visual acuity the first problem is to determine how the physical parameters of this image are related to those of the external scenes that give rise to it. We will assume that these scenes produce incoherent light, since that is the normal situation in natural vision. However, coherent light enters the discussion when we consider experiments that use interference techniques to create very high frequency retinal images. (The *Appendix* to this chapter, p. 302, describes the difference between coherent and incoherent imagery.)

RETINAL IMAGE REPRESENTATIONS OF OBJECT PROPERTIES. *Shape and size.* The strongest refractive surface in the eye is the anterior surface of the cornea, this air-to-cornea interface accounting for about two-thirds of the overall refractive power of the eye. Essentially all of the remaining refraction occurs at the surfaces and within the internal stria of the crystalline lens. The interacting effects of these various refractive elements are complex, but for a normal eye and a

distant object, a good first-order approximation to their optical behavior can be made by substituting for all those elements a single thin lens with a focal length of 17 mm, located 17 mm in front of the retina, or about 7.5 mm behind the anterior surface of the cornea (45, 84, 144). The center of this equivalent lens is called the nodal point of the simplified eye. To find the location in the retinal image of any given point in a scene, one draws a straight line from the point through the nodal point of the eye, and the intersection of that line with the retina is the location of the corresponding region in the retinal image. Because the distance from the nodal point of the eye to the retina is about 17 mm, the size of the image of any object in the plane perpendicular to the line of sight, or as projected onto that plane, is given by

$$\frac{S_i}{S_o} = \frac{17}{d_o}$$

where S_o is the size of the object, or its projection on the perpendicular plane, and d_o is the distance from the nodal point of the eye to the object, in mm.

Object and retinal image sizes are often expressed in units of visual angle, that is, the angle subtended by the object or its image at the nodal point of the eye. For example, an object 1 m long, in a plane perpendicular to the line joining one of its ends with the nodal point of the eye at a distance of 100 m, subtends a visual angle of

$$\tan^{-1}\frac{1}{100} = 0.57^\circ = 34'$$

Intensity relations. When an object that is radiating light is viewed, the intensity at each point in its image, that is, the number of quanta per unit time per unit area incident on the retina in the region corresponding to the object, is directly proportional to *a*) the intensity of light radiated by the object in the direction of the eye, and *b*) the area of the pupil of the eye. As the light travels from the objects to the retina it passes through various media, e.g., air, corneal tissue, etc., that absorb, reflect, or scatter some proportion of it, reducing the intensity of the image. If pupil size and the fraction absorbed, scattered, and reflected are fixed, the intensity of the retinal image is linearly related to the intensity of the light radiating from the object. Thus the retinal image contains information about an important physical parameter of the object, namely how much light it radiates in the direction of the eye. However, the reliability of that information depends upon the extent to which the pupil has a constant area. Because the normal pupil ranges in diameter from about 8 mm to 2 mm, the intensity of the retinal image of a given luminous object can vary over a range of about 16:1. Consequently, unless a brain mechanism knows the size of the pupil (information not directly available to introspection, for example), it cannot know this parameter exactly, but only within an interval of roughly one log unit. The information contained in the retinal image about the relationships among the intensities of two or more sources seen simultaneously is much better, because pupil size affects all their images by the same factor.

In the natural environment in which the eye evolved most objects do not actively generate light, but instead, reflect light generated elsewhere. The intensity of the retinal image of a reflective object depends jointly on a property of the object, reflectance, and the intensity of incident light. The retinal images of reflective objects thus do not contain information specifying their reflectance directly, but only the product of incident illumination and reflectance. Because the reflectances of ordinary objects vary from about 90% to about 10%, that is, over a range of only about 10:1, while the incident illumination varies by more than 10^6:1 in our normal environment, the intensity of the retinal image of an object does not uniquely tell us the reflectance of the object. However, the retinal image does contain accurate information about the relative reflectances of two or more objects if it is known that they are equally illuminated. Over a broad range of light intensities (luminance > 1 cd/m^2) the increment-threshold intensity is proportional to the background intensity (i.e., $\Delta I/I$ = constant—Weber's law.) Consequently over this range reflectance differences that are visible at any illumination level will be visible at all levels.

Optical effects that blur the retinal image. When the object viewed is a point source, that is, a source of light that subtends a negligibly small visual angle (a star, for example), several properties of the eye will cause light from the source to be spread over some finite area of the retina.

Focus errors. Figure 1 represents a simplified eye looking at a point source. Some of the rays from the source are refracted by the equivalent lens, pass through the pupil, and strike the retina.

The rays in this figure are shown crossing in front of the retina, representing a condition where the image of the point source falls in front of the retina as it would in an uncorrected myopic (nearsighted) eye. If the pupil of the eye is circular, then the retinal light distribution predicted by geometrical optics will be a uniform disk whose diameter increases linearly with the distance between the retina and the point where the rays cross (that is, the focus error), and also linearly with the diameter of the pupil. This disk is called a "blur circle." (The shape of the distribution

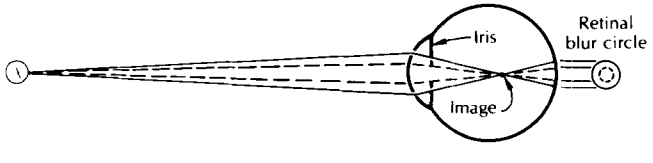

FIG. 1. Blur circles in a schematic misfocused eye for two different pupil sizes. [From Cornsweet (36).]

will be the same as the shape of the pupil. If the pupil were square, it would be a "blur square.") Blur circles for pupils of two different diameters are shown in Figure 1.

If two or more stars are imaged at the same time, each will form its own blur circle, and, because light intensities add linearly, the resulting retinal light distribution will simply be the sum of the individual blur circles. An extended source can be considered as an array of independent point sources, each of which forms its own blur circle. The resulting retinal light distribution is simply the sum of the distributions from each of these points. Since the blur circle for each point is identical, the defocused image of any object is simply the light distribution in the object convolved with the appropriate blur circle. [The operation of convolution is described mathematically in the *Appendix* to this chapter, *Shift-Invariant Linear Operators and Convolution*, p. 303 and TWO-DIMENSIONAL CASE, p. 304. Intuitively, this amounts to a smearing process in which each point in the object is imaged as a disk of light on the retina; the entire image of the subject is the sum of these (normally overlapping) disks.]

Aberrations. Even if a point source is focused as well as possible on the retina, the light from the source will still be distributed over a finite area as a result of imperfections or aberrations in the refractive system of the eye.

Spherical aberration (a consequence of differences in refractive power across the pupil as a function of distance from its center) varies strongly from one eye to another, within a given eye as a result of changes in accommodation, and from one meridian to another in the same eye.

Chromatic aberration results from differences in refractive power as a function of the wavelength of the incident light. When middle wavelength rays from a heterochromatic stimulus are in focus on the retina, rays from the blue end of the spectrum come to a focus in front of the retina and rays from the red end converge "behind" the retina. The magnitude of this aberration is relatively constant across individuals, and is quite substantial, as shown in Figure 2. Note that when the eye is accommodated for the red end of the spectrum, which seems to be the natural state of accommodation in viewing distant targets [(84), p. 45–46] deep blue light from the same target will be roughly 2 diopters out of focus.

Astigmatism (resulting from differences in refractive power as a function of meridional angle, i.e., for an eye with a horizontal line of sight, the power of the eye in a vertical plane might be different from its power in a horizontal plane) causes the retinal light distribution from each point in an object to be more or less elliptical in shape, the lengths of the major and minor axes of the ellipse depending upon the amount of astigmatism and the corresponding focus errors. Eyes differ greatly both in their amount of astigmatism and in its meridional orientation.

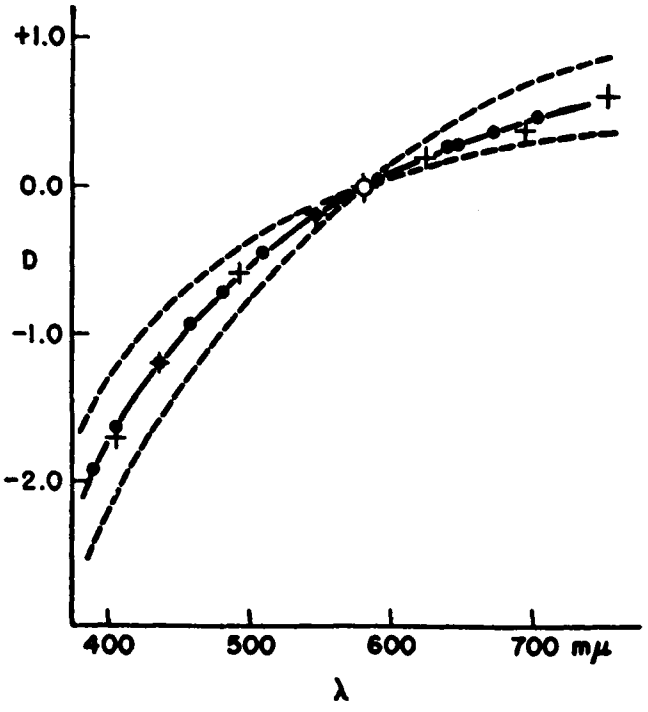

FIG. 2. Chromatic aberration of the human eye. λ, Wavelength; D, magnitude of chromatic aberration in diopters (assuming the eye is in perfect focus for 578 nm, D gives the power of the spectacle lens required to focus other wavelengths). *Solid line* and *solid dots*, average from 12 observers of Bedford and Wyszecki (19); *crosses*, average from 14 observers of Wald and Griffin (133); *dashed lines*, total range for all observations of Bedford and Wyszecki. [Adapted from Bedford and Wyszecki (19).]

Irregular refractive strengths over the surfaces of the eye, as in Figure 3, further degrade the image of each point in the object. The pattern of these irregularities varies considerably across observers and, for a given observer, across states of accommodation. Figure 3 shows a map of "isopower" lines across the pupil of one observer at zero accommodation; at higher levels of accommodation the map for this observer looks very different (125).

Coma refers to complex aberrations that arise as the angle between a collimated beam and the lens is varied. This effect has recently been shown to contribute importantly to the overall aberration of the eye, and to vary among different eyes (72).

Diffraction. Even if none of the aberrations just discussed were present in the eye, the light from a point source would spread over a finite region of the retina as a consequence of diffraction at the pupil. This can be thought of as an inevitable consequence of the fact that an optical instrument of limited size can only intercept a portion of the wave front arising from a source. The resulting distribution is approximately Gaussian, with a diameter that varies inversely with pupil diameter and directly with wavelength. (See Figs. 6 and 7.)

Scattering within the retina. As noted earlier the retina has an optically appreciable thickness: before

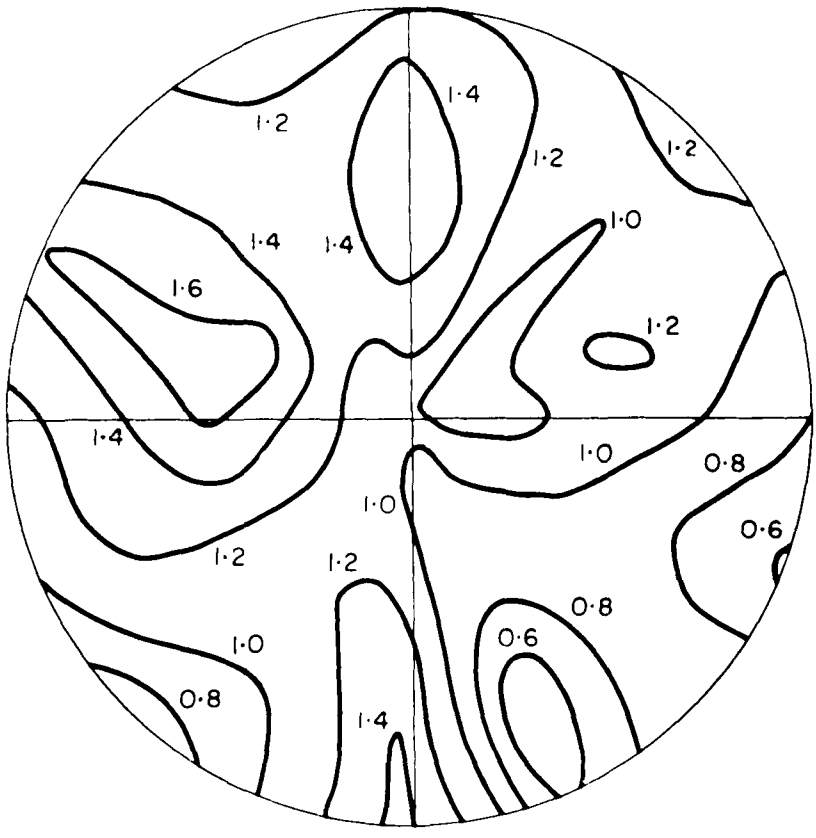

FIG. 3. Irregularities in refractive power across the pupil of a normal eye during relaxed accommodation (pupil diam, 7.2 mm). Contours join points of equal excess power (diopters). [From van den Brink (29a), with permission from Pergamon Press, Ltd.]

reaching the outer segments, light must pass through a layer of neural tissue that ranges from roughly 80 μm at the center of the fovea to nearly 400 μm at 5° eccentricity. This tissue and its vascular system must scatter some light, and that has generally been regarded as the functional explanation for its thinness in the region specialized for high acuity [but cf. Rodieck (106), p. 368]. Light that passes through the outer segment layer must also scatter upon reflection from the underlying tissues. Measurements of retinal scatter have recently been reported by Gorrand (52).

Pointspread and linespread functions. As a result of all of these factors, the light distribution in the focused retinal image of each point in an object is spread over a region of the retinal surface. The two-dimensional function that describes the image distribution for a single object point located at the origin is called the *pointspread function*: Figure 4 illustrates the concept for an idealized optical system consisting of a single lens. If the pointspread function is known for a given eye, the actual light distribution in the retinal image of any scene can be obtained by convolving the light distribution in the scene with the pointspread function. (In actuality the pointspread function varies with distance from the center of the fovea, and so in principle the entire set of pointspread functions must be known in order to derive the retinal image. However, because the high acuity phenomena that concern us here are mediated by the fovea, a single pointspread function is sufficient for present purposes.)

In practice, direct measurement of the pointspread function of the human eye is very difficult. (In fact, it has not yet been achieved.) But this function can be calculated from measurements of the *linespread function*, which are somewhat easier to obtain. The linespread function describes the cross section of the image of a thin (in principle, infinitely thin) line, as illustrated in Figure 5: If this function is known, and the optical system is circularly symmetrical (so that the linespread function does not depend on the orientation of the object line), the pointspread function can be derived from it by Fourier analytic methods, as outlined in the *Appendix* to this chapter, p. 302.

Figure 6 shows the foveal linespread function of the human eye for various pupil sizes [as measured by Campbell and Gubisch (33)], and Figure 7 shows the profiles of the corresponding pointspread functions as calculated by Gubisch (59). (In this instance it happens that the linespread and pointspread functions look roughly the same, but mathematically this need not always be the case.) These figures show two spread functions for each pupil size. The narrower function in each case is the theoretical spread that would result from diffraction at the pupil alone, while the wider function is the actual observed line or point spread. Notice that as pupil size increases the effect of diffraction at the pupil decreases but the actual spread

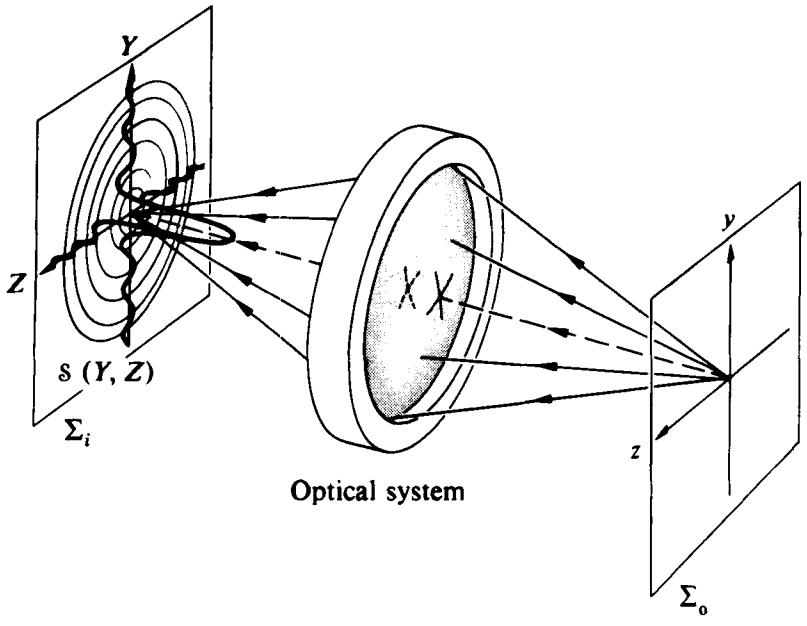

FIG. 4. Pointspread function of an optical system. Σ_o is the object plane, with coordinates y,z; Σ_i, is the image plane; $S(y,z)$, irradiance distribution produced in Σ_i by point source located at origin of Σ_o. Illustration is schematic. [From Hecht and Zajac (62).]

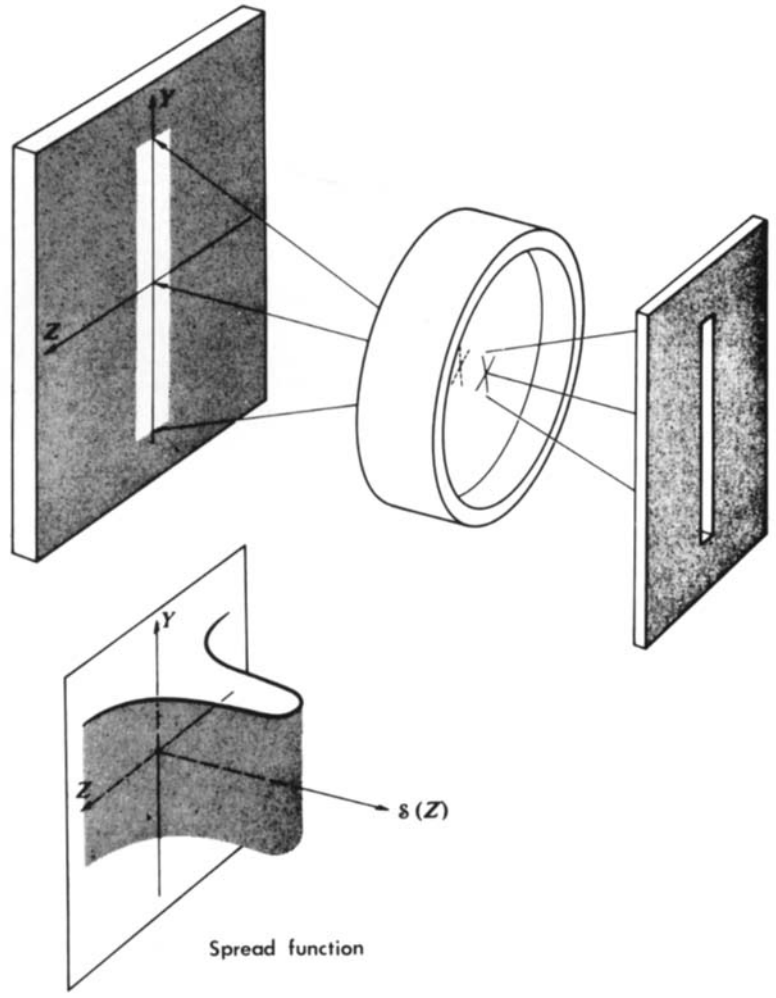

FIG. 5. Linespread function of an optical system. $S(Z)$ is the irradiance distribution in the image plane produced by an infinitely long thin line source in object plane. (The absence of secondary ripples here like those in Fig. 4 has no significance; both illustrations are schematic.) [From Hecht and Zajac (62).]

increases a bit due to the increased importance of other factors. For pupil diameters ≤2 mm (the lower limit in natural viewing) the observed spread functions are practically the same as those that would result from diffraction alone in an idealized optical system. (The light source here was a broad-band white light with a spectrum approximating the photopic luminosity function. Diffraction predictions were made by calculating monochromatic diffraction patterns for many different wavelengths and then weighting and averaging the resulting curves.)

The modulation transfer function. If the pointspread function of an optical system is known, the light distribution in the image of any object can be derived by convolving the light distribution in the object with the pointspread function. As mentioned above, the pointspread function can be derived from measures of the linespread function. An alternative procedure is to measure the *modulation transfer function* (MTF).

To measure the MTF one uses as an object a sinusoidal grating, that is, a pattern for which intensity in one direction varies sinusoidally, while intensity is uniform in the orthogonal direction, as shown in Figure 8. The *contrast C* (or *modulation*) of such a grating is defined as

$$C = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

where I_{max} is the maximum intensity, I_{min} the minimum.

The image of such a grating will also have a sinusoidal intensity distribution, regardless of the nature of the pointspread function. (The convolution of a sine wave with any other function remains sinusoidal.) The ratio of image to object contrast is a measure of how well the optical system transfers the modulation of the object to the image plane. The fidelity of this transfer, that is, the value of this ratio, plotted as a function of the spatial frequency (i.e., cycles/deg) of the object grating, is the MTF of the optical system. (Figure 9 shows the MTF of a typical human eye for three different pupil sizes [as calculated by Gubisch (59) on the basis of the linespread data in Fig. 6]. For this optical system, and for all ordinary optical systems,

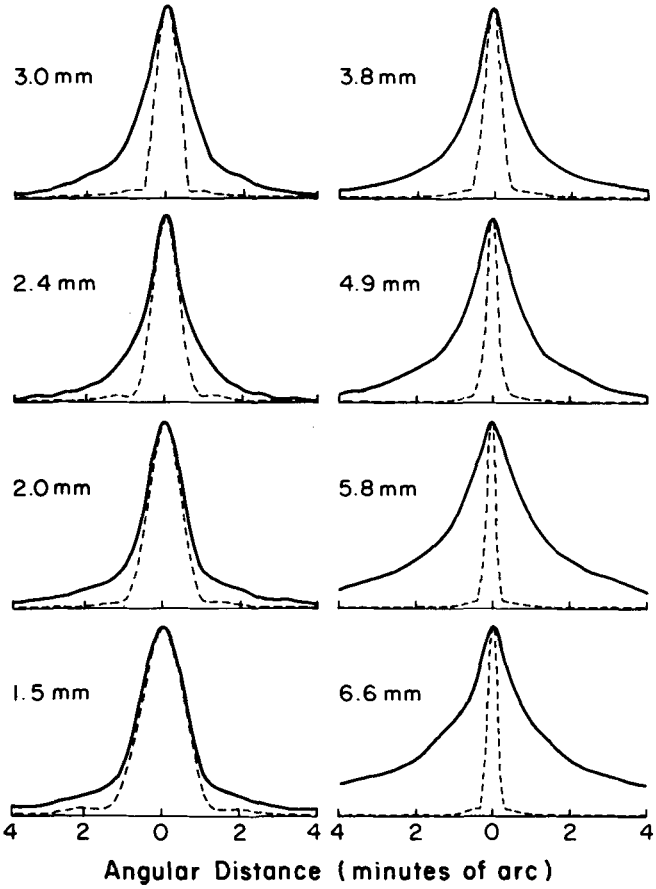

FIG. 6. Linespread functions of the human eye for white light at various pupil diameters. *Heavy dotted curves*, measured intensity (along horizontal axis) of retinal image of a line target as measured ophthalmoscopically. *Narrower dashed curves*, theoretical spread resulting from diffraction at the pupil alone. (All curves have been normalized to equal 1.0 at origin.) [From Campbell and Gubisch (33), with permission from Cambridge University Press.]

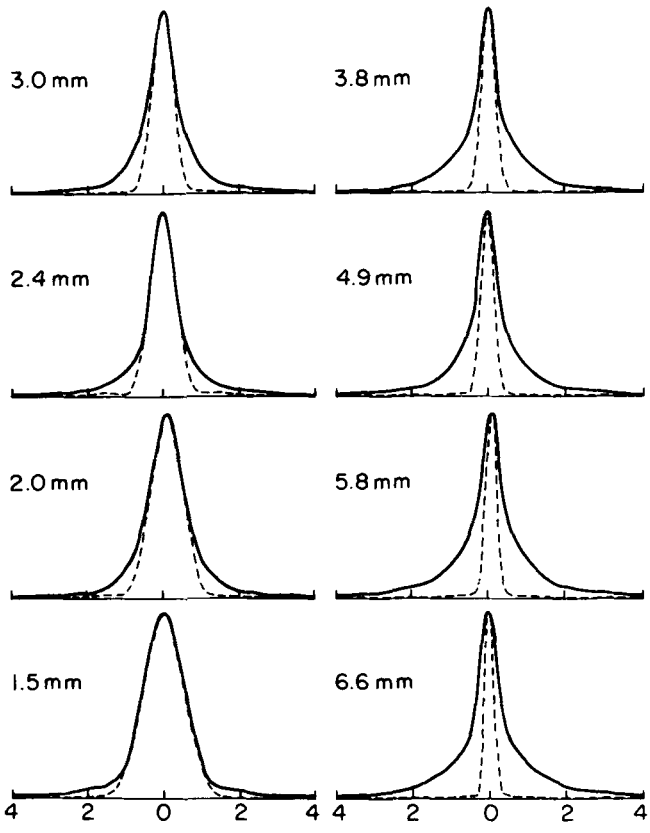

FIG. 7. Pointspread functions of the human eye for white light at various pupil diameters. *Solid curves*, radial profiles of theoretical pointspread functions calculated from linespread curves in Fig. 6. *Narrower dashed curves*, the point spread resulting from diffraction alone. [From Gubisch (59).]

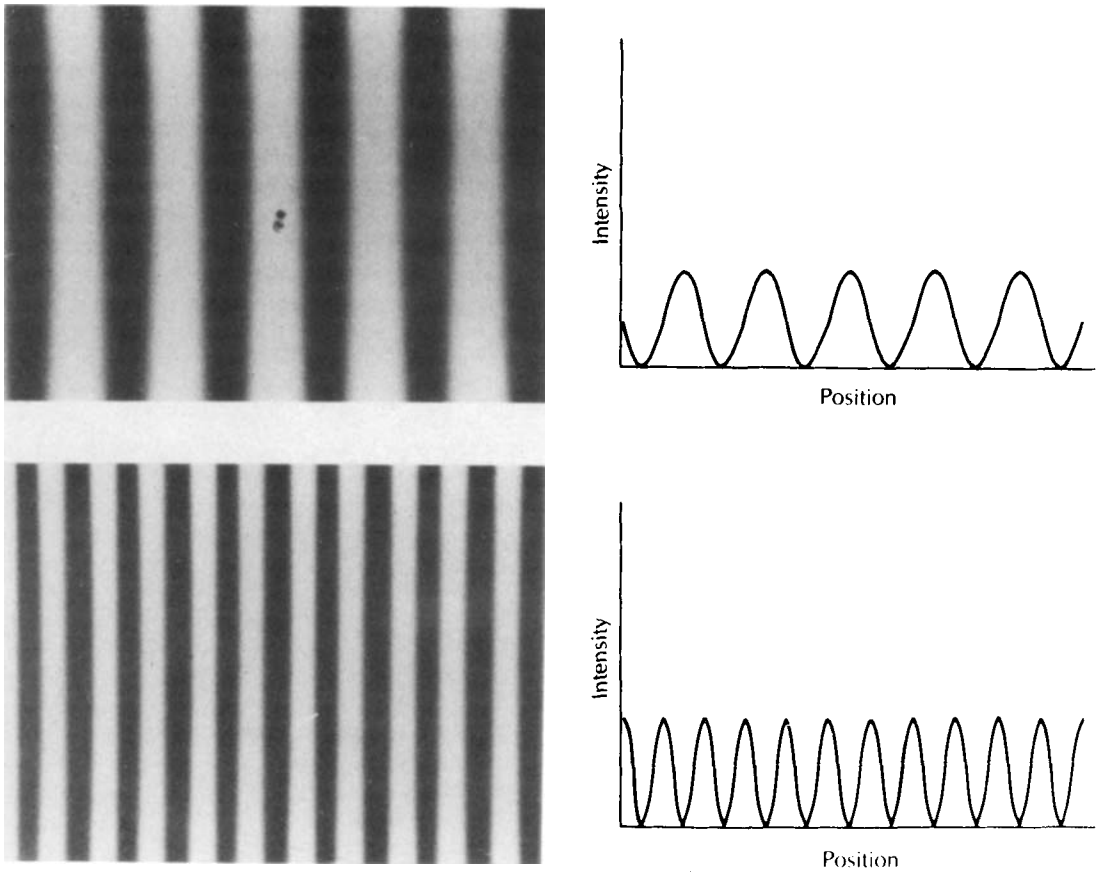

FIG. 8. Sinusoidal gratings. [From Cornsweet (36).]

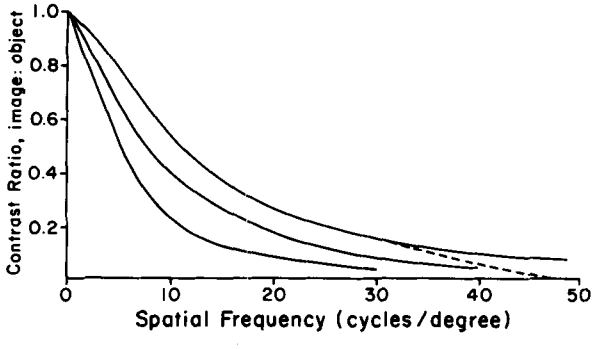

FIG. 9. Modulation transfer functions (MTFs) of the human eye for various pupil diameters. Each curve shows contrast reduction (i.e., retinal image contrast divided by the object contrast) imposed on sinusoidal targets as function of their spatial frequency. *Solid curves* (from *top* to *bottom*) correspond to pupil diameters of 2.4, 3.8, and 6.6 mm; *dashed curve* corresponds to 1.5 mm. MTFs calculated from linespread data of Fig. 6. [From Gubisch (59).]

the contrast ratio approaches one as the spatial frequency approaches zero. That is, for low spatial frequencies, the image contrast is as great as the object contrast, and as spatial frequency increases, the contrast in the image decreases relative to the object, until at very high frequencies (that is, for very closely spaced gratings) the image contrast is essentially zero regardless of the contrast of the object. Once the MTF of an optical system is known its linespread and pointspread functions can be calculated by Fourier analytic methods. The mathematical details of this calculation are spelled out in the *Appendix* to this chapter, p. 302, but the essential fact is that the MTF is the absolute value (i.e., modulus) of the one-dimensional Fourier transform of the linespread function, or equivalently, the profile of the absolute value of the two-dimensional Fourier transform of the pointspread function. Before taking absolute values here, one has what is called the *optical transfer function* (OTF). When the OTF is entirely real and nonnegative—as will be the case when the optical system produces no phase shifts—the MTF and OTF are identical, and so the distinction need not be preserved. Both conditions are met by a well-focused eye, and in the vision literature one usually encounters only the MTF. However, it is important to keep in mind that the MTF by definition must be nonnegative and consequently does not reflect effects in which the OTF becomes negative for some spatial frequencies. Such effects can be produced by defocusing as discussed later in *Defocus and spurious resolution*, p. 268.

In general, then, if the MTF of a lens is known, the light distribution in the image of any object can be derived by *1*) analysis of the object into its Fourier spectrum, *2*) multiplication of the spectrum with the

MTF (or more precisely with the OTF, in cases where the MTF and OTF are not identical), and 3) Fourier synthesis of the resulting product spectrum. Because of the directness of this operation the quality of an image-forming system is usually specified in terms of its MTF rather than its point- or linespread function.

Defocus and spurious resolution. As noted earlier, defocusing can be modeled by assuming that the image is convolved with a blur circle whose width is directly proportional to the distance between the retina and the actual location of the image plane. In the frequency domain this convolution corresponds to multiplication of the transfer function by the Fourier transform of a disk (Fig. A2 in the *Appendix* to this chapter, p. 302, illustrates this transform). Broadly speaking, the effect of the multiplication is to reduce contrast in the image, with greater reductions for higher frequencies. However, at a finer level of analysis (spelled out in the *Appendix* to this chapter, p. 302) it turns out that the actual effect will be to reduce contrast first to zero (i.e., as frequency increases); then to *reverse* the contrast of the image for higher frequencies—over this range absolute contrast first increases a bit, and then declines again to zero; next to re-reverse the contrast for the next range of frequencies, and so on. Thus a sinusoidal input with spatial frequency in one of the reversed contrast ranges will appear in the image 180 deg out of phase, i.e., with its peaks turned into valleys, and vice versa. In the optical literature, this phenomenon is known as "spurious resolution."

Figure 10 illustrates this process in terms of its effect on the MTF, and compares the theoretical optical results (dashed lines) with the actual effect of defocus on human contrast sensitivity as determined by Campbell and Green (32). (The dashed lines in the figure represent the MTF, which it will be recalled is the absolute value of the OTF, hence nonnegative. The contrast reversal regions correspond to intervals between the first and second zeros of the MTF, also between the third and fourth, etc. This illustrates the point that the MTF by itself is sometimes misleading, since it can only specify amplitude reductions, and not contrast reversals, which from a technical standpoint are phase-shift effects, the sinusoidal input being shifted by half a period in the output.)

The reader can observe these spurious resolution effects subjectively by carefully examining the upper spoke pattern in Figure 11 with one eye from a distance close enough that it is blurred on the retina. (The lower pattern is an out-of-focus photograph of the upper one, designed to illustrate how the latter will appear when viewed very close up.) The apparent contrast decreases toward the center of the figure, where spatial frequency is highest. Looking carefully, one can see a narrow ring or hourglass-like region where the contrast is zero. Inside that ring, the stripes are again visible, but with *reversed* contrast (a consequence of the convolution referred to above), then another ring and another contrast reversal, etc. If the observer has no astigmatism, the ring of zero contrast

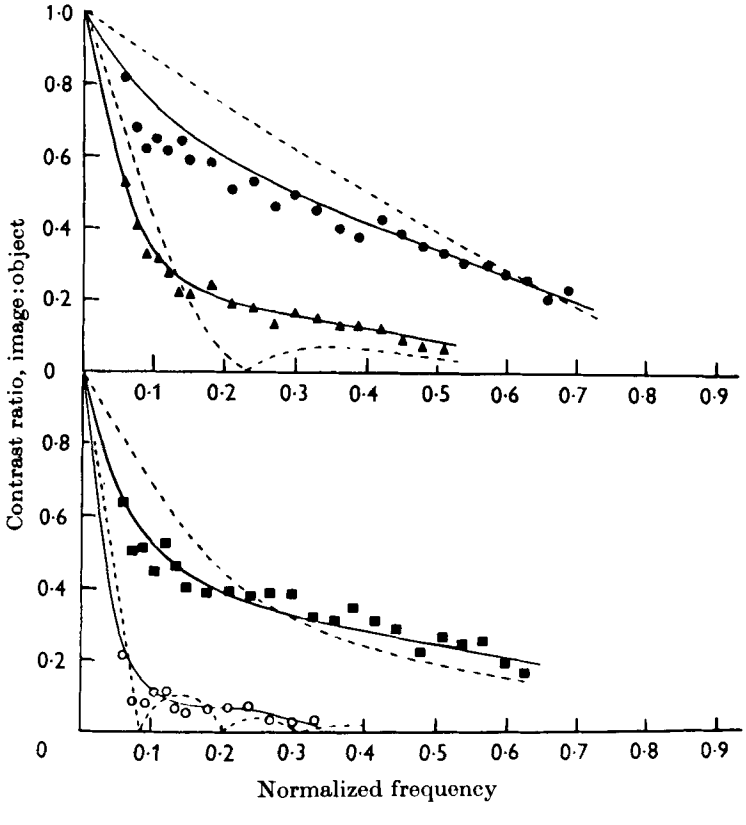

FIG. 10. Effects of defocus on human spatial contrast sensitivity (2-mm pupil). Data points show retinal image contrast reduction (inferred from threshold data of one observer) imposed on each spatial frequency as a function of defocus. *Upper panel*: ●, 1.5 diopters (myopic); ▲, 2.5 diopters. *Lower panel*: ■, 2.0 diopters, ○, 3.5 diopters. Spatial frequency expressed as fraction of highest frequency passed by the optics of the eye, i.e., 1.0 ≈ 60 cycles/deg. *Solid lines* drawn by hand through data points. *Dashed lines* show results predicted from diffraction alone. [From Campbell and Green (32), with permission from Cambridge University Press.]

will be round. If it is hourglass-shaped, the long axis of the hourglass is the meridian that yields the largest blur circle, and the orthogonal meridian will probably yield the smallest blur circle and thus the small axis of the hourglass. (The reader may also notice that when this spoke target is viewed from a distance at which it can be sharply resolved, the central region when fixated appears tinged with yellow. This is not due to macular screening pigment, because fixation on a blank portion of the page does not produce the same effect. We suspect this illusion may be related to the low density of blue cones in the central fovea, as discussed later in THE OPTICAL TRANSFER FUNCTION AND PHOTORECEPTOR SPACING, p. 270.)

Spatial Modulation Detection

HIGH-FREQUENCY RESOLUTION UNDER NORMAL VIEWING CONDITIONS. We have seen that for a number of physical reasons the retinal image is normally a degraded copy of the visual stimulus. The combined effect of all these degrading factors can be summarized by the MTF of the eye (or, more precisely, its OTF), which specifies the attenuation imposed on each spatial frequency in the stimulus and consequently allows us to determine how any stimulus will appear at the level of the retina. Clearly the visual system cannot detect stimulus information that is not physically present in the retinal image, and so the modulation transfer function sets an upper bound on the ability to resolve fine spatial detail under normal viewing conditions. ("Normal" here means incoherent illumination viewed with a natural pupil. Under special conditions, described below, it is possible to bypass the optics of the eye and form retinal images containing spatial frequencies higher than the normal cutoff.)

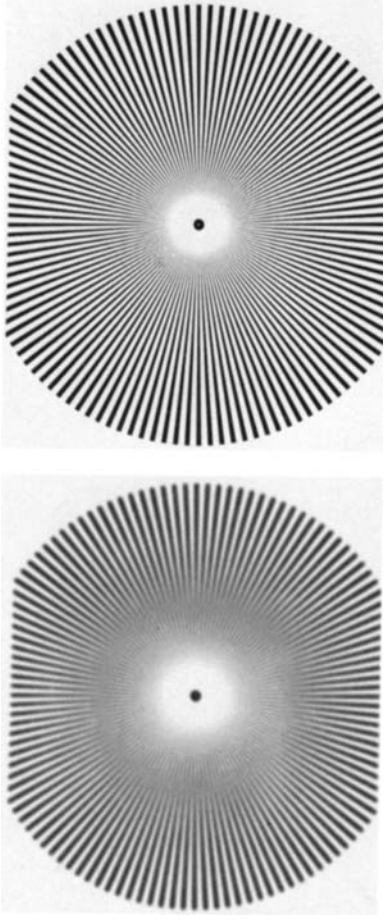

FIG. 11. *Top*: spoke target for demonstrating spurious resolution in human vision. See text for directions. *Bottom*: out-of-focus picture of spoke pattern illustrating spurious resolution in a photographic image.

Suppose in particular that the stimulus is a vertical sinusoidal grating (as in Fig. 8) with intensity profile

$$B[1 + C\cos(2\pi\phi x)]$$

where x is measured in units of visual angle along the horizontal axis of the retina (i.e., degrees or minutes), ϕ is the spatial frequency (cycles/unit of visual angle), B is the mean intensity, and C is the contrast of the sinusoidal modulation. ($0 \leq C \leq 1$). At the level of the retina the image of this stimulus will take the form

$$aB[1 + t(\phi)C\cos(2\pi\phi x)]$$

where a is an attenuation factor representing light loss due to reflection and preretinal absorption in the ocular media, and $t(\phi)$ is the OTF of the eye evaluated at frequency ϕ. Consequently if we require an observer to discriminate between one stimulus in which $C = 0$ (i.e., a uniform field of intensity B) and another in which C is some nonzero value (in the extreme, $C = 1$) we are really asking him to discriminate between the retinal images aB and $aB[1 + t(\phi)C\cos(2\pi\phi x)]$. Obviously discrimination is impossible when $t(\phi) = 0$. The MTFs of Campbell and Gubisch [Fig. 9; (33)] show that for the human eye under optimal conditions, $t(\phi)$ vanishes around 50–60 cycles/deg. We can conclude from this that spatial contrast in the stimulus at frequencies higher than 1 cycle/min can never be detected because, in effect, it never reaches the retina: under normal viewing conditions this represents the physically imposed upper limit of visual acuity.

Of course it would be entirely possible for human contrast sensitivity to be substantially worse than the ideal limits imposed by optical considerations, and in fact this is the case under low light conditions and—at all intensity levels—for stimuli presented outside the fovea. However, under optimal conditions the performance of the visual system as a whole matches the upper limit imposed by the optics of the eye. Figure 12 shows psychophysical contrast thresholds for a variety of mean intensity levels as determined by van Ness and Bouman (126). These curves plot the smallest value of C at which an observer can discriminate $B[1 + C\cos(2\pi\phi x)]$ from a uniform field of intensity B. (Such plots are known as "spatial contrast sensitivity functions." Sometimes they are loosely referred to

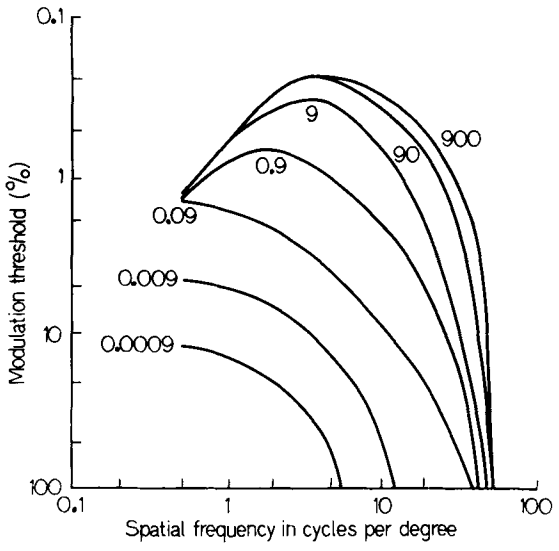

FIG. 12. Human spatial contrast thresholds for vertical sinusoidal gratings at various mean intensity levels (2-mm pupil, monochromatic light, wavelength 525 nm). *Each curve* plots threshold contrast as a function of spatial frequency for mean retinal illumination level (trolands) indicated by curve parameter (100% corresponds to contrast of 1.0). [Data from Van Ness and Bouman (126); figure from Westheimer (140).]

as "modulation transfer functions," by analogy to actual MTFs as determined for lenses and other imaging devices.) At the highest intensity levels (B = 90 and 900 trolands) contrast thresholds become independent of mean intensity (Weber's law), and discrimination becomes impossible when ϕ reaches approximately 55 cycles/deg (trolands = target luminance in candelas/m^2 × pupil area in mm^2). This is in good agreement with the upper limit of 50–60 cycles/deg predicted from the optical MTF. Many other psychophysical measurements have led to the same results (184). The data in Figure 12 were obtained with vertical gratings. Horizontal gratings give rise to similar values, but sensitivity to oblique gratings is somewhat worse. This difference disappears beyond a degree of retinal eccentricity ranging from 8° to 18° (20) and apparently is due to central effects, since it shows up in cortical evoked responses but not in the electroretinogram (91).

THE OPTICAL TRANSFER FUNCTION AND PHOTORECEPTOR SPACING. The fact that the normal retinal image never contains spatial frequencies higher than 1 cycle/min has implications for the optimal spacing of photoreceptors. Clearly, if the retina consisted simply of one giant receptor which integrated quantum catch over the entire retinal surface, the eye would be incapable of discriminating spatial modulation at all frequencies—i.e., it would be limited to registering only the overall mean intensity of the retinal image. On the other hand there would be no evident utility to packing in more individual receptors than is justified by the quality of optics of the eye. Fourier analytic considerations show that the optimal center-to-center spacing between receptors should equal $(2\phi_c)^{-1}$, where ϕ_c is the highest spatial frequency that can be present in the retinal image. (This is the well-known "Sampling Theorem": its mathematical basis is spelled out in the *Appendix* to this chapter, p. 302). At this spacing the receptor mosaic is theoretically capable of providing an undistorted reproduction of any retinal image, i.e., no harmonic distortion will be introduced by the fact that the continuous optical image is sampled by an array of discrete receptors. Taking ϕ_c to be 1 cycle/min, this analysis implies an optimal center-to-center spacing of 0.5 minute. Osterberg's (95) cell counts of the human retina (Fig. 13) and the data of Polyak (98) indicate a peak cone density of roughly 150,000–200,000 cells/mm^2 at the center of the fovea; this corresponds to a spacing of 0.59–0.51 minute. Evidently in the central fovea there is a good match between the dimensions of the receptor mosaic and the optics of the eye: Any coarser sampling would lose (or distort) information potentially available in the retinal image, while any finer sampling would be superfluous, since it would never reveal any new information.

The idea that overall receptor density in the center of the retina has evolved to match the spatial frequency cutoff imposed by the optics of the eye raises the question of how this density should be parceled out among the three spectrally distinct classes of cones. At present there is no direct anatomical evidence as to the actual spacing of "red," "green," and "blue" cones in the human retina, because in primates there is no known way of identifying a cone's spectral class by its appearance alone. However, staining tech-

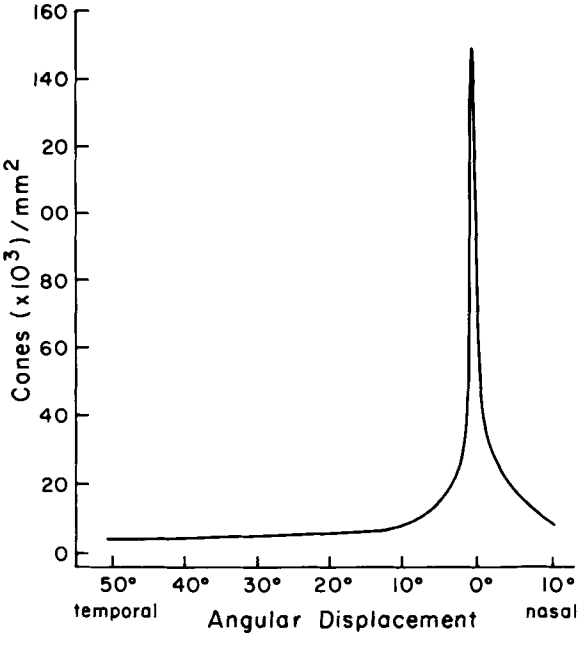

FIG. 13. Cone density as a function of distance from the center (0°) of the fovea. [Data from Osterberg (95); figure adapted from Ripps and Weale (103).]

niques that allow such an identification have recently been introduced and successfully applied to the baboon retina [see Marc and Sperling (94)]. Consequently it seems likely that definitive information on human cone spacing will be available before long, and so some speculation seems in order. Here we concentrate on the blue cones, whose spacing has long been a subject of controversy [(29), p. 240–244].

Clearly the key optical factor here is chromatic aberration, which as we have seen is quite substantial in the human eye (Fig. 2). Apparently when accommodation is entirely relaxed the eye is normally in focus for the red end of the spectrum [Ivanoff's data for 10 observers, cited in Le Grand (84), suggests a range from 625 nm to 725 nm]. Then with increasing accommodation the focal wavelength shifts downwards (sparing the eye some accommodative effort), reaching a lower limit around 500 nm for targets requiring 2.5 diopters of accommodation (40-cm viewing distance). Consequently retinal images in the range 400–500 nm are always out of focus by an amount ranging from 0 diopters to 2 diopters of myopia. This means that over the spectral range to which the blue cones are most sensitive the OTF of the eye goes to zero not at 60 cycles/deg but at some lower frequency, the exact value of which depends on pupil size and the assumed magnitude of "blue myopia." Sampling considerations in turn suggest that the density of blue cones ought to be lower than that of red and green cones.

To make an exact prediction of blue cone spacing based on the sampling theorem one needs to assume that the visual system is wired for some "typical" combination of pupil size and amount of blue myopia. Then one can calculate the spatial frequency cutoff of the corresponding optical transfer function—call this spatial frequency ϕ_b—and predict the blue cone spacing $s_b = (2\phi_b)^{-1}$. [Formulas for the OTF of a misfocused retinal image are given in the *Appendix*, Equations A17 and A18. As noted earlier in our discussion of spurious resolution (see *Defocus and spurious resolution*, p. 268) the OTF here goes monotonically to zero as frequency increases, and thereafter oscillates—within a monotonically shrinking envelope—between negative and positive values. For present purposes, it seems natural to identify the cutoff frequency ϕ_b with the first zero of the OTF.] The difficulty is that it is not obvious what pupil size and degree of blue myopia to assume. One plausible course is to assume the smallest normal pupil (2 mm diam) and 1–2 diopters of myopia—as though the retina were designed for distant viewing in bright light, with a focal wavelength of 675 nm [the average for Ivanoff's observers (in ref. 84)], and the blue cones spaced to sample an image in the range 400–500 nm. Then 1 diopter of blue myopia yields a cutoff frequency $\phi_b = 10.2$ cycles/deg and a spacing $s_b = 3'$ visual angle, while 2 diopters yields $\phi_b = 4.8$ cycles/deg, with $s_b = 6.3'$.

The later figure agrees closely with Marc and Sperling's (94) measurement of blue cone spacing in the center (±0.5°) of the baboon fovea (mean = 6′), where overall cone density is roughly the same as in the human retina. However, in baboon, blue cone density is not maximal at the center of the fovea (as it is for both red and green cones) but instead at an eccentricity of 1°, where the spacing averages 3′—exactly the prediction for 1 diopter of blue myopia.

For the spacing of blue cones in the human retina we have only psychophysical evidence—based either on spatial mappings of detection thresholds for blue points superimposed on yellow backgrounds designed to silence the red and green cones (132, 141) or on visual acuity for targets visible only to blue cones (29, 41, 57). Both methods indicate that blue cone spacing in man has the same characteristics as in the baboon: The density is low in the very center of the fovea, with no evidence of any blue cones in a region subtending somewhere between 8′ (132) and 20′ (141). (Assuming a 6′-intercone distance we would expect only 2 blue cones in the first area and 11 in the second. Thus there is no serious discrepancy between these psychophysical observations of a "blue-blind" central zone and the notion that the spacing is actually constant at 6′ out to an eccentricity of ±0.5°). Outside this tiny central region the mapping technique (141) has produced an estimate of 10′ spacing out to 40′ from the center of the fovea, while acuity measurements with foveal targets yield spatial frequency cutoffs ranging from 9 ± 1 cycles/deg (41, 47) down to 4 cycles/deg (29). These cutoffs translate into blue cone spacings ranging from 3.33′ up to 7.5′—figures very close to the upper and lower spacings found in baboon fovea. [These estimates assume that the psychophysically determined spatial frequency cutoff is related to blue cone density by the sampling theorem. There is no necessary reason why this should be so, and it has been conjectured—e.g., by Brindley (29), that the poor acuity of the blue system is due to convergence of many blue cones onto the same second-order neuron rather than a sparsity of the cones themselves. However the weight of current evidence runs counter to that hypothesis.]

In summary then, what we know about blue cone spacing lends support to the idea that not only is overall foveal cone density matched to the overall spatial frequency cutoff of the retinal image, but in addition the same principle governs the individual densities of the various color systems. In a nutshell, blue cones are relatively sparse because from their spectral perspective the retinal image is always relatively coarse. This idea in turn raises intriguing questions as to how the visual system integrates spatial information arising from receptor systems that have different sampling capabilities. Speculation on this topic, however, would carry us too far afield.

BYPASSING THE OPTICS OF THE EYE. Using coherent light (e.g., from a laser source) it is possible to create interference patterns on the retina that contain spatial

frequencies well beyond the normal 1-cycle/min cutoff. The technique, introduced in the 1930s by Le Grand (84), involves forming on the retina a diffraction pattern equivalent to the one produced by a pair of small apertures illuminated by the same coherent source. (In this case the eye acts like an optical computer of Fourier transforms as noted in the *Appendix* to this chapter, p. 302.) This procedure allows one to bypass the normal optical limitations of the eye (except for scattering within the retina itself), and determine whether the neural components of the visual system are capable of resolving spatial frequencies higher than those normally present in the retinal image. Figure 14 shows contrast thresholds as a function of spatial frequency measured under these conditions by Campbell and Green (32). The basic finding is that here, just as in normal viewing, resolution becomes impossible at spatial frequencies on the order of 1 cycle/min. (This was also found by Le Grand (84), Westheimer (138), and Green (58). The only exceptional results are those of Byram (31), who obtained limits on the order of 2.5 cycles/min. These results are discussed later in *Neural Limits of Visual Acuity*, p. 275.) Figure 15 shows a comparison between contrast thresholds obtained with interference patterns and normal targets at the same mean retinal illuminance: In both cases the high-frequency portions of the curves are fairly well fitted by straight lines (in this semilog plot) with the line for normal viewing having a somewhat steeper slope—necessarily, since in this case the effective contrast at the retina is attenuated by the MTF of the optics of the eye, which does not affect the contrast of the interference targets.

Why should the postoptical components of the visual system be incapable of resolving spatial modulation beyond 1 cycle/min? Clearly from the standpoint of design efficiency such a limitation makes sense, since under normal viewing conditions no higher frequencies would ever be informative about the outside world—i.e., they could only arise as a result of optical phenomena within the eye itself. However, the question still arises as to the actual mechanisms responsible for the loss. Two kinds of factors need to be considered: first, purely physical limitations imposed by the size and spacing of the photoreceptors and by residual optical spreading that affects interference targets as well as normal ones; and second, physiological limitations due to neural mechanisms.

Physical factors affecting resolution of interference targets. Dimensions of receptor array. Anatomical evidence (95, 98, 104) indicates that in the central fovea [what Polyak (98) called the "foveola," a circular region 80′ across] the cone outer segments form a dense array of tightly packed cylinders, each having a cross section shaped somewhere between a circle and a hexagon—so that seen end-on "the entire formation here resembles an evenly distributed mosaic, like the cobblestones in an old fashioned pavement" [(98), p. 269; see Fig. 18 below). Receptor width in this region is approximately 0.25 min, the length of the outer

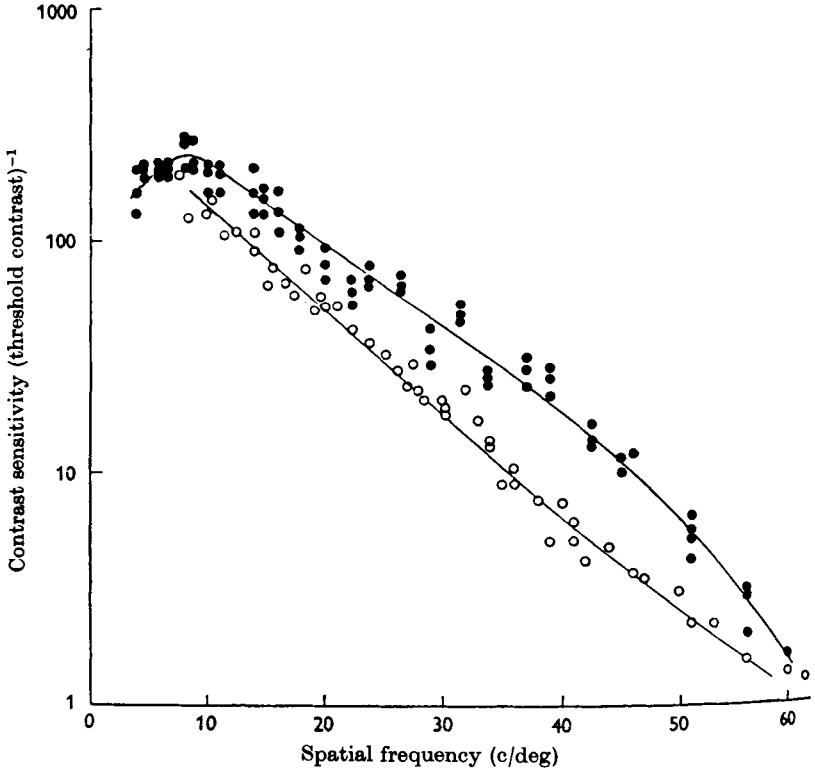

FIG. 14. Spatial contrast-sensitivity functions for sinusoidal interference targets formed with a coherent source; wavelength, 633 nm. Data for two observers; smooth curves drawn by hand through data points. Mean illumination level, 500 trolands. [From Campbell and Green (32), with permission from Cambridge University Press.]

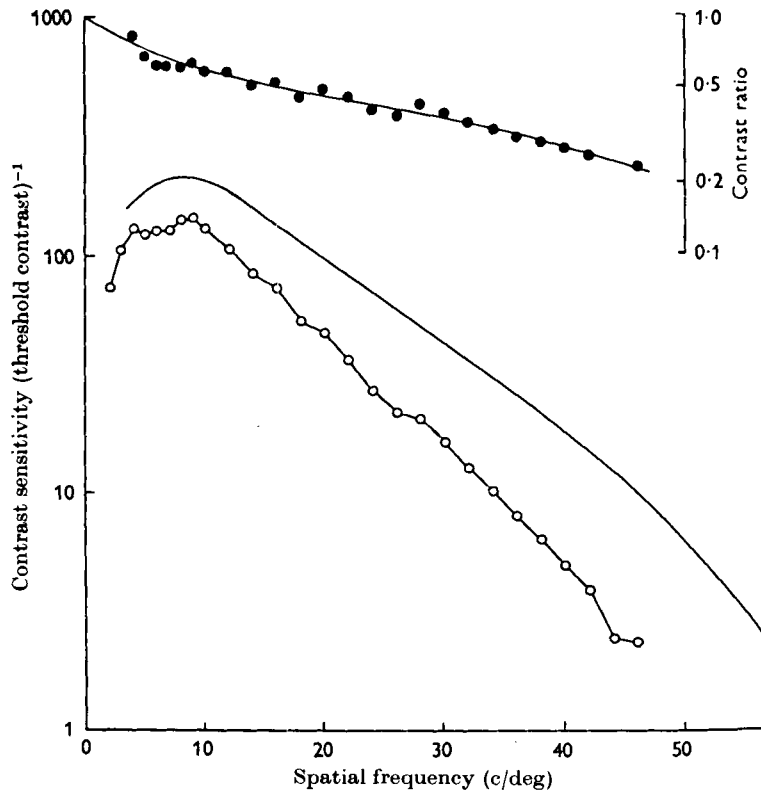

FIG. 15. Spatial contrast-sensitivity functions for interference fringes and normal targets at same mean retinal illuminance (500 trolands). Continuous smooth curve in *lower portion* of figure taken from data (*closed circles*) of Fig. 14 (i.e., these are interference target thresholds). *Open circles* show contrast thresholds for targets viewed normally, i.e., through the optics of the eye. *Closed circles* in the *top portion* show the ratio between the two curves below: this is an inferred measure of the modulation transfer function of the optics of this observer's eye. [From Campbell and Green (32), with permission from Cambridge University Press.]

segments is roughly 30 times their width, and the center-to-center distance between receptors is approximately 0.5 min. The foveola contains approximately 10^4 cones and no rods. As one moves out from the central fovea, cone density falls off approximately as the inverse distance from the center (as shown in Fig. 13), and the cones themselves become shorter and fatter. Conversely, rod density increases, reaching a peak approximately 20° from the center. (Rod dimensions are not of primary importance here because the upper limits of visual acuity are always obtained at mean illumination levels high enough to guarantee rod saturation.)

Residual optical factors. In terms of these dimensions it would not take a great deal of optical spreading within the retina to produce an effective spatial resolution cutoff of 1 cycle/min. As an example (motivated by the roughly linear falloff in semilog coordinates shown in Fig. 14) an optical spread corresponding to the MTF e^{-ms} (where s denotes spatial frequency in cycles/min) would reduce contrast by 95% (i.e., from 1.0 to 0.05) at 1 cycle/min. The corresponding point-spread function in this case (which is given explicitly by Eq. A17 in the *Appendix*, p. 302) would be roughly Gaussian, with a height of 0.64 at the origin and 0.05 one min away. Thus light would be spread over a circular region having an effective radius of 3 receptors. In other words, the 1-cycle/min cutoff obtained with interference targets could be produced by a relatively small spread of light within the retina itself.

The actual behavior of light within the retina is far from fully understood. (Ref. 117 surveys the relatively new field of photoreceptor optics, which studies the waveguide properties of receptors.) However, Gorrand (52) has recently reported a few measurements of retinal scattering based on a new ophthalmoscopic technique that allows one to decompose the linespread function into two separate components corresponding to preretinal spread and scattering within the retina. (The technique used to produce the linespreads in Fig. 6 lumps both together.) Psychophysical methods can also be used to separate preretinal and intraretinal spread (32): The top portion of Figure 15 represents a psychophysical estimate of the preretinal MTF. Both physical and psychophysical methods yield about the same value for preretinal spread in the fovea (52), and agree in showing hardly any intraretinal spread in that region (where the retina is, of course, very thin). Consequently it seems fairly certain that optical spreading is not the major factor limiting foveal acuity for interference patterns—though in view of the relatively large effect to be expected from even a very small spread, and the fact that only a few physical measurements have so far been reported, it cannot be concluded that it plays no role at all.

Outside the fovea Gorrand's (52) measurements show that retinal scatter becomes more appreciable. For example at 6° eccentricity the contrast of a 7-cycle/deg interference grating is reduced by a factor around 0.7, and that of an 11-cycle/deg grating by 0.4.

However, here it is even more certain that retinal spreading is not the major factor limiting acuity, because at 6° the highest resolvable frequency for interference patterns is 6–8 cycles/deg (see ref. 58; Fig. 16). In other words a 7-cycle/deg grating that has an effective contrast at the outer segments of 0.7 is just at the threshold of visibility, whereas at the center of the retina the same frequency can be detected when its contrast is less than 0.005 (Fig. 12). Evidently postoptical differences between these two regions of the retina must be responsible for an effective contrast reduction of more than two log units.

Receptor sampling effects. The first postoptical factor that differentiates various regions of the retina, is of course, receptor density (Fig. 13). We have already noted that in the center of the fovea, receptor density matches the value prescribed by the sampling theorem for reconstructing inputs up to 60 cycles/deg, and this is also the psychophysical acuity limit for that region. If receptor density is the critical factor limiting visual acuity it is natural to expect that local acuity should be systemically related to local receptor density. Green (58) has measured local acuity for interference gratings in 30-min patches out to 8° eccentricity and compared the results to local receptor densities based on Osterberg's (95) cell counts. Figure 16 shows his results. It can be seen that out to 2° there is a striking agreement

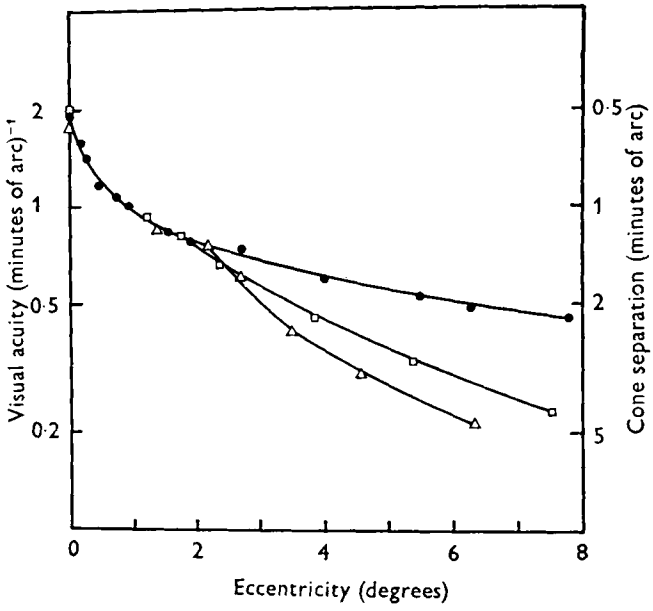

FIG. 16. Comparison between local spatial acuity for interference targets (633 nm) and intercone spacing at various distances from the center of the fovea (0° eccentricity). *Closed circles* show center-to-center intercone distances (1/distance in min). *Open points* show local spatial acuity for two observers. Visual acuity here is arbitrarily defined as twice the highest resolvable spatial frequency in cycles/min. This definition allows a direct comparison between local spatial acuity and local cone density: if acuity and cone density were matched according to the sampling theorem, both should fall on a common curve—as they do out to 2°. Targets subtended 32.4 min; mean intensity level was 1,200 trolands. [From Green (58), with permission from Cambridge University Press.]

between local spatial acuity (i.e., highest resolvable spatial frequency) and local interreceptor spacing: Letting s_d denote the highest resolvable frequency at distance d from the center of fovea, and w_d the center-to-center receptor spacing at distance d, Figure 16 shows that out to $d = 2°$

$$s_d = (2w_d)^{-1}$$

As Green pointed out, this relationship has a natural interpretation in terms of information theory. As we noted earlier, Fourier analytic considerations (the Whittaker-Shannon sampling theorem, described in the *Appendix*, p. 302) show that a two-dimensional array of sample points spaced w units apart (e.g., an array of photocells located at the nodes of a checkerboard with square size $w \times w$) permits perfect reconstruction of any image in which the highest spatial frequency does not exceed $(2w)^{-1}$. On this basis one can say that from 0° to 2° local spatial resolution capacity equals "the theoretical limits for a mosaic of receptors" [Green (58)].

However, this neat relationship should not be interpreted as an explanation of local spatial acuity, because the sampling theorem does not imply that spatial frequencies above $(2w)^{-1}$ simply vanish when sampled by an array with interreceptor distance w—in the sense that their sampled output will be indistinguishable from that produced by a uniform field. Rather, $(2w)^{-1}$ only represents the highest frequency that can be resolved without distortion: If a regular sampling array is confronted with frequencies above this limit it *can* transmit them, but their sampled output will be identical to that produced by lower frequencies. In communication theory this sort of distortion is known as "aliasing" (27, 96). [The mathematical details are explained in the *Appendix* to this chapter, p. 302, which also provides a figure illustrating the effect (Fig. A4).] To avoid aliasing in artificial image transmission systems one must prefilter the input to ensure that the sampling array never receives spatial frequencies higher than $(2w)^{-1}$: Mathematically this is accomplished by convolving the input with a pointspread function whose MTF vanishes at $(2w)^{-1}$. (Ideally this MTF should be 1.0 at all lower frequencies for optimum efficiency, but this usually cannot be achieved in practice because the corresponding point spread must contain negative regions.) In the human eye, of course, such prefiltering is normally performed by the optical transfer function, which vanishes at 1 cycle/min—the appropriate cutoff for receptor spacing in the central fovea. However, outside this region the optically imposed frequency cutoff will not be appropriate for the coarser sampling dimensions of the receptor mosaic (and of course with interference targets it does not operate at all, except for residual spreading within the retina itself). Consequently on sampling grounds alone there is no reason why spatial frequencies above $(2w)^{-1}$ should be entirely invisible in a retinal interference pattern—one might just as reasonably expect that

substantially higher frequencies could be detected, though their appearance would be that of low-frequency counterparts produced by aliasing.

In summary then, residual optical spreading within the retina cannot account for the limits of visual acuity for interference targets: In the center of the fovea it is apparently physically negligible, and outside the center its effects, while physically appreciable, are nowhere near large enough to explain the dramatic decline in acuity. Out to 2° local resolution capacity matches the $(2w)^{-1}$ sampling theorem limit, but this relationship is itself more of a puzzle than an explanation of visual acuity, because sampling considerations alone do not imply that higher spatial frequencies should be invisible: The $(2w)^{-1}$ limit only specifies the range of frequencies that can be transmitted without distortion, and some additional mechanism is necessary to explain why higher frequencies are apparently undetectable. Beyond 2° spatial resolution capacity is substantially worse than the sampling theorem cutoff. Finally, it should be noted that local spatial acuity in scotopic (rod-mediated) vision is best around 4° from the center of the fovea, while rod density is greatest around 20° (124). Thus it appears that receptor geography alone cannot account for any spatial acuity limits. Consequently in order to explain the overall spatial resolution capacity of the visual system one has to turn to neural mechanisms.

Neural Limits of Visual Acuity

We have seen that everywhere on the retina the highest detectable spatial frequency either equals or falls below the aliasing cutoff implied by local receptor density. The problem now is to understand how these limits are produced by neural mechanisms. In the recent past, spatial contrast detection has been the subject of a great deal of theoretical work aimed at constructing physiologically realistic models that can account for psychophysical phenomena in terms of the receptive-field properties of cells in the visual pathway (53–55, 142). This kind of neural model building has reached quite a high level of sophistication, and one might imagine that even though many details remain to be worked out, at least the general neural principles underlying spatial contrast sensitivity are well understood—in the same way that we understand the principles underlying trichromacy.

However, if one begins to trace the flow of spatial contrast information up the visual pathway, beginning with the distribution of quantum catch in the receptor mosaic, one immediately encounters a theoretical road block that does not seem to be dealt with by any of the current models. This section focuses on this difficulty, which we shall call the "aliasing problem." The conclusion of our analysis is that because of this problem it is not at all clear that current theories of spatial contrast detection actually capture the fundamental principles underlying the neural mechanisms that limit visual acuity. To make this point we first give a brief overview of current theoretical ideas, and then explain why something more seems necessary to account for the limits of visual acuity.

PSYCHOPHYSICAL POINTSPREAD MODELS. In his review of theories of visual acuity, Le Grand [(84), p. 102] distinguished a class of "continuous theories" which "voluntarily neglect the discontinuous receptive structure of the retina." Current models for spatial contrast sensitivity (that is, models designed to account for the spatial contrast sensitivity function and similar data) are continuous theories, and we shall argue that as a consequence they cannot give a complete account of visual acuity. All of the models we have in mind are motivated primarily by the well-known antagonistic center-surround receptive-field organization first identified in retinal ganglion cells, and now known to characterize cells at every level of the visual pathway from receptors (16–18, 47) to level IV of the striate cortex (68). The discovery of this general organizational feature of the visual system corroborated Ernst Mach's (in ref. 101a) nineteenth century prediction—based entirely on perceptual analysis—that the light distribution in the retinal image must be spatially filtered by neural processes of lateral inhibition and excitation. To model this process it is natural to imagine that the neural components of the visual system operate on the retinal image in a fashion analogous to an optical pointspread function, but with the important difference that optical pointspread functions must be nonnegative, whereas a neural point spread can be both positive and negative. (This is critical, because a nonnegative point spread necessarily produces an MTF that is maximal at the origin and consequently could not explain the nonmonotonic shape of human spatial contrast sensitivity functions obtained at moderate-to-high mean luminance levels—e.g., the curves above 0.09 troland in Fig. 12.) If this neural filtering were linear its effect could be directly combined with the point spread produced by the optics of the eye to yield an overall pointspread function which, convolved with any stimulus, would yield its "neural image." The contrast in this neural image would then determine the limits of spatial contrast sensitivity: if the neural contrast for a given spatial frequency is zero when its physical contrast is 1.0, the frequency must be undetectable.

Figure 17*A* and *B* illustrate this kind of model in its basic form. Figure 17*A* shows spatial contrast sensitivity functions for several observers obtained in an experiment by Kelly and Savoie (76) that concentrated on the low-frequency end of the spatial spectrum. Also shown is a theoretical curve of the form s^2e^{-s} which provides a good fit for all observers. Suppose now that spatial contrast detection obeys the following model: the stimulus pattern S is first convolved with a pointspread function P (which combines both optical and neural spread) and the result $P*S$ is Fourier transformed to obtain its amplitude spectrum. If for any

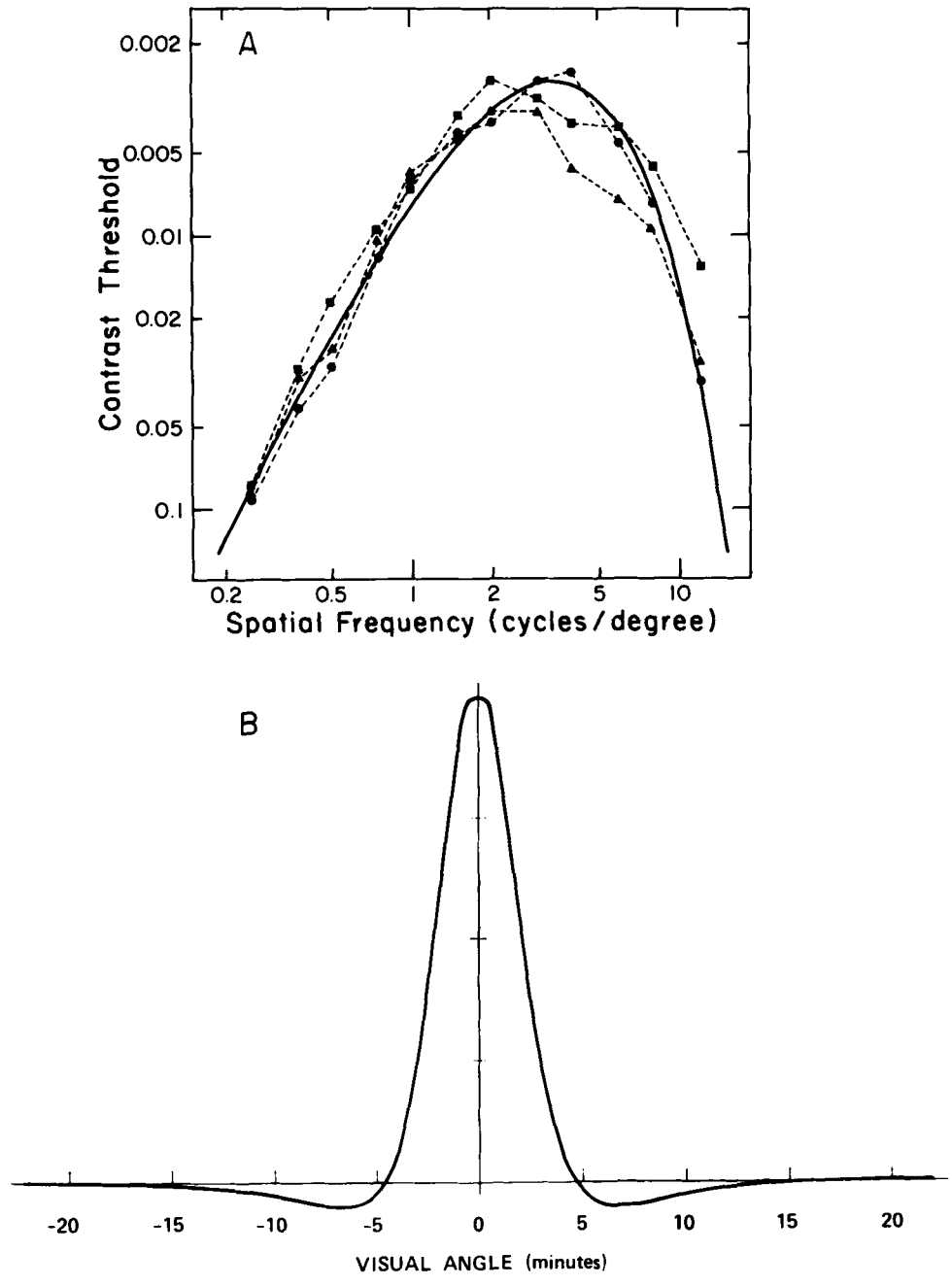

FIG. 17. *A*: spatial contrast-sensitivity functions for three observers (*dashed lines*) with fitted function s^2e^{-s} (*solid curve*), which can be regarded as the overall "modulation transfer function" of the human visual system. *B*: pointspread function obtained by Fourier inversion of the smooth curve in Figure 17*A*. [From Kelly (74).]

spatial frequency the amplitude spectrum exceeds a fixed threshold t, the observer detects contrast. When the stimulus is a sinusoidal grating of the form $1 + C \cos 2\pi sx$, where s is spatial frequency in cycles/unit of x (e.g., cycles/deg) and C is contrast, then detection will occur when $C|T_p(s)| = t$, where $|T_p(s)|$ is the absolute value of the Fourier transform of the pointspread function P. In other words, $|T_p(s)|$ is the modulation transfer function of this model visual system. Thus in this model the threshold contrast $C(s)$ for detecting a sinusoid of frequency s is inversely proportional to $|T_p(s)|$, and so the spatial contrast sensitivity function $(C(s))^{-1}$ (e.g., as plotted in Fig. 17*A*) is directly proportional to the modulation transfer function of the visual system. Consequently to obtain the psychophysical pointspread function P (up to a constant factor reflecting the threshold t) one can simply calculate the inverse Fourier transform of the spatial contrast sensitivity function. Figure 17*B* (74) shows the outcome of applying this principle to the contrast

sensitivity function (i.e., the fitted curve) in Figure 17*A*. The result, according to this model, should be the overall pointspread function of the visual system, incorporating both optical and neural effects. The function in Figure 17*B* is obviously reminiscent of the center-surround organization of the receptive fields of retinal ganglion cells, and so it is tempting to suppose that spatial contrast sensitivity can be explained physiologically by a combination of optical blurring and lateral neural mechanisms of a kind that are well established.

This kind of model was first employed in connection with sinusoidal contrast detection by Schade (115), who had in mind the practical task of incorporating human contrast sensitivity into the design of television systems. In effect, the human visual system could in this fashion be treated as the final stage in a series of linear filtering operations. Subsequent psychophysical work (see ref. 66, chapt. 1) coupled with results from neurophysiology (ref. 66, chapt. 2) has led to a considerable elaboration of the basic idea. Current models (36a, 53, 142) are designed to take account of the inhomogeneity of the visual field (by allowing the psychophysical pointspread function to change shape as a function of retinal position), and of the possibility that each point on the retina is served by several "mechanisms" or "channels," each having its own point spread (hence, its own spatial bandwidth) and its own sensitivity to the temporal parameters of the stimulus (i.e., some channels prefer flashes, others slow changes). For example, the recent model of Wilson and Bergen (142) assumes four channels at each point on the retina.

These psychophysical pointspread models account for the limits of local visual acuity in terms of the effective contrast transmitted by the most sensitive local channel, i.e., the channel which imposes the smallest contrast reduction on high spatial frequencies. If the output of this channel is below threshold for a grating of physical contrast 1.0, the limit of acuity has been reached. (Actually the detection process is currently assumed to involve "probability summation" between channels, so that the less-sensitive channels still contribute something to detection even when other, better tuned, channels are present. This makes a difference in estimating channel bandwidths, but is irrelevant to our main point here.)

Now there is no question that the most highly evolved of these psychophysical spread models can make good predictions of spatial contrast sensitivity functions, and of contrast thresholds for certain other stimuli besides sinusoidal gratings. Why then do we say that they are not obviously capable of accounting for the limit of visual acuity, which after all is only the high-frequency cutoff of the contrast sensitivity function? Put more broadly, could not lateral neural processes act in a manner analogous to optical point spread and thereby limit visual acuity by reducing the effective contrast of high-frequency retinal images?

The answer, in a nutshell, is that by neglecting the discontinuous receptive structure of the retina (i.e., the receptor sampling process) these models bypass the aliasing problem mentioned earlier, BYPASSING THE OPTICS OF THE EYE, Receptor sampling effects, p. 274. The key point is that spatial frequencies above the aliasing cutoff implied by local receptor density must be represented in the *input* to the neural part of the visual system as counterfeit low frequencies and therefore cannot be filtered out by the kind of static pointspread operations envisioned by current models. The next section explains this point in detail.

SPATIAL AMBIGUITY OF THE RECEPTOR IMAGE: THE ALIASING PROBLEM. Suppose that in a given patch of retina the receptors are laid out in a checkerboard array (i.e., each centered on an intersection), with center-to-center interreceptor distance w deg, and receptor width pw ($0 < p \leq 1$). Then suppose we image on this patch a vertical sinusoidal grating of frequency f cycles/deg and contrast C (i.e., a grating of the form $1 + C \cos 2\pi fx$). The aliasing cutoff in our patch is $(2w)^{-1}$: If f is always below this limit we know that in principle it can be reconstructed from its sampled image. However, if $f > (2w)^{-1}$ it can be shown that its sampled image will be identical to that produced by a lower frequency sinusoid of the form $1 + C \cos 2\pi(w^{-1} - f)x$, i.e., a sinusoid of frequency $w^{-1} - f$ and contrast C'. (This is worked out in the *Appendix* to this chapter, p. 302: see Eq. A23. The contrast reduction from C to C' is due entirely to integration over the surface area of the receptor). For example, if $w = 0.008°$ (0.5 min) and $pw = 0.004°$, as in the center of the fovea, the aliasing cutoff $(2w)^{-1}$ is 60 cycles/deg, and a 90-cycle/deg grating of contrast C will produce the same sampled image as a 30-cycle/deg grating of contrast $0.6C$. Similary, if $w = 0.033°$ (2 min) and $pw = 0.008$ (to simulate the dimensions at 8° eccentricity) the aliasing cutoff is 15 cycles/deg and a 25-cycle/deg grating of contrast C produces the same sampled image as a 5-cycle grating of contrast $0.8C$.

Thus as a consequence of the aliasing possibilities inherent in receptor sampling there is an intrinsic spatial ambiguity in the input to the visual system: Any given distribution of quantum catch in the receptor mosaic—say a distribution consistent with spatial frequency f—could either have been produced by a real stimulus having frequency f, or by a higher frequency (here, $w^{-1} - f$) masquerading as f. However, the psychophysical literature tells us that frequencies greater than $(2w)^{-1}$ are seen as uniform fields, while lower frequencies are seen as spatially modulated. This means that the neural component of the visual system somehow manages to solve the aliasing problem without prefiltering the retinal image—a feat television engineers might well envy. How is it possible?

To forestall potential misunderstandings and objections, four points should be noted immediately. First, it should be clear that under the assumptions we began

with (the realism of which we discuss in a moment) no neural mechanism of any kind can discriminate between the *stationary* retinal images of real spatial frequencies and their aliases (e.g., between 30 cycles/deg and 90 cycles/deg when w = 0.5 min) This is so because after sampling by the receptors these images present identical inputs to any subsequent neural process.

Second, it should not be thought that aliasing is only a real possibility with unnatural stimuli such as interference patterns. That is true in the center of the fovea, where the spatial limits of the eye match the aliasing cutoff of the receptors, but outside that small region the retina must be routinely exposed to spatial frequencies above its local aliasing cutoffs, because the optical quality of the retinal image remains roughly constant out to 30° (84) while receptor density decreases dramatically. (The example for 8° eccentricity given earlier illustrates this point.) Consequently the aliasing problem is one the visual system encounters in its normal operation, and presumably must be designed to solve.

Third, it is natural to wonder whether aliasing is only a mathematical possibility that depends critically on the unrealistic assumption of precisely regular receptor spacing. Certainly the real geometry of the receptor mosaic is not nearly as regular as that of our checkerboard model, and the exact consequences of irregular receptor spacing pose an open theoretical problem. However, one can get some immediate insight by constructing sampling arrays that simulate the actual distribution of receptors in real retinas. We have used Polyak's photographs of the human retina [(98), p. 268] to construct arrays of pinholes that simulate the sampling effects of the central fovea (by simply poking a pin through the apparent center of each receptor). When such an array is superimposed on gratings simulating frequencies beyond 60 cycles/deg (e.g., displayed on a television screen) one sees aliasing effects in the form of moire patterns that look like low-frequency "worms"—broad, elongated, irregular shapes that change as the grating moves (Fig. 18 illustrates the effect). This shows that the irregularities of actual receptor spacing do not automatically eliminate aliasing effects. Rather, it seems that they make these effects less exact: instead of a frequency beyond the $(2w)^{-1}$ cutoff aliasing back into a single lower frequency (e.g., 90 cycles/deg aliasing into 30 cycles/deg), it looks as though the effect of irregularity is to scatter a given high frequency into a range of lower frequencies, giving rise to patterns comparable to low-frequency spatial noise. However, there is no obvious reason why these low-frequency patterns should not be detected by the visual system and used to discriminate very high frequency gratings from uniform fields. In other words, the realities of receptor geography do not seem likely to provide a solution to the aliasing problem.

Finally, if aliasing is a real visual phenomenon (i.e., a real problem of the visual system, rather than simply a theoretical possibility), why has it not been revealed in previous experiments, i.e., in form of acuity values higher than the aliasing cutoff? (Of course, in a sense these would be artificial acuities mediated by a kind of false resolution, because the frequencies actually seen would not be physically present in the retinal image. However, that is not the issue.) Perhaps it has, because the acuity literature is too enormous for a thorough review, but the fact that one does not find results of this sort reported routinely could very well be due to three points of methodology. To demonstrate super acuity mediated by aliasing one must first ensure that frequencies higher than the local aliasing cutoff actually reach the retina. Consequently measurements in the fovea made with ordinary stimuli could never be expected to yield acuities higher than 60 cycles/deg. Second, it is essential to control image motion (i.e., by flashing the stimulus, or stabilizing its retinal image) because when motion is allowed the visual system has available information that in principle could allow it to discriminate between real spatial frequencies and counterfeits due to aliasing (e.g., if the eye moves at v deg/s a real 30-cycle/deg grating produces 30 v Hz flicker at each receptor, while a 90-cycles/deg grating masquerading as 30 cycles/deg will produce 90 v Hz flicker). As far as we know all of the experiments measuring acuity with interference targets have allowed sustained viewing with natural eye movements. Third, because detection of aliased frequencies would be a somewhat unnatural task, it seems essential to use an objective psychophysical method, e.g., forced choice discrimination between uniform and spatially modulated fields. This point is underscored by the exceptional results of Byram (31) mentioned earlier. He reported acuity limits as high as 150 cycles/deg, but stressed that the choice of an upper limit depended on what one was willing to accept as a legitimate experience of spatial contrast: At his highest values, regular square-wave gratings appeared as "wriggling curved line segments" [(31), p. 273]. Because Byram gave few details of his methods, and all other interference experiments have produced acuity values approximating 60 cycles/deg, his results may have been due to some artifact. However, it is also possible that he happened to achieve conditions facilitating actual detection of aliased frequencies.

Outside the fovea aliasing could be expected to show itself with ordinary stimuli, provided they were flashed or presented as stabilized images and objective psychophysical methods were used. We are not aware of any experiments satisfying both requirements. However, Le Grand (84) mentions the interesting point that in the periphery of the retina, unlike the fovea, acuity is significantly better for flashed targets than for steady ones. It seems conceivable that this is due to false resolution mediated by aliasing, though other explanations based on the greater prevalence of transient ganglion cells in the periphery are also plausible.

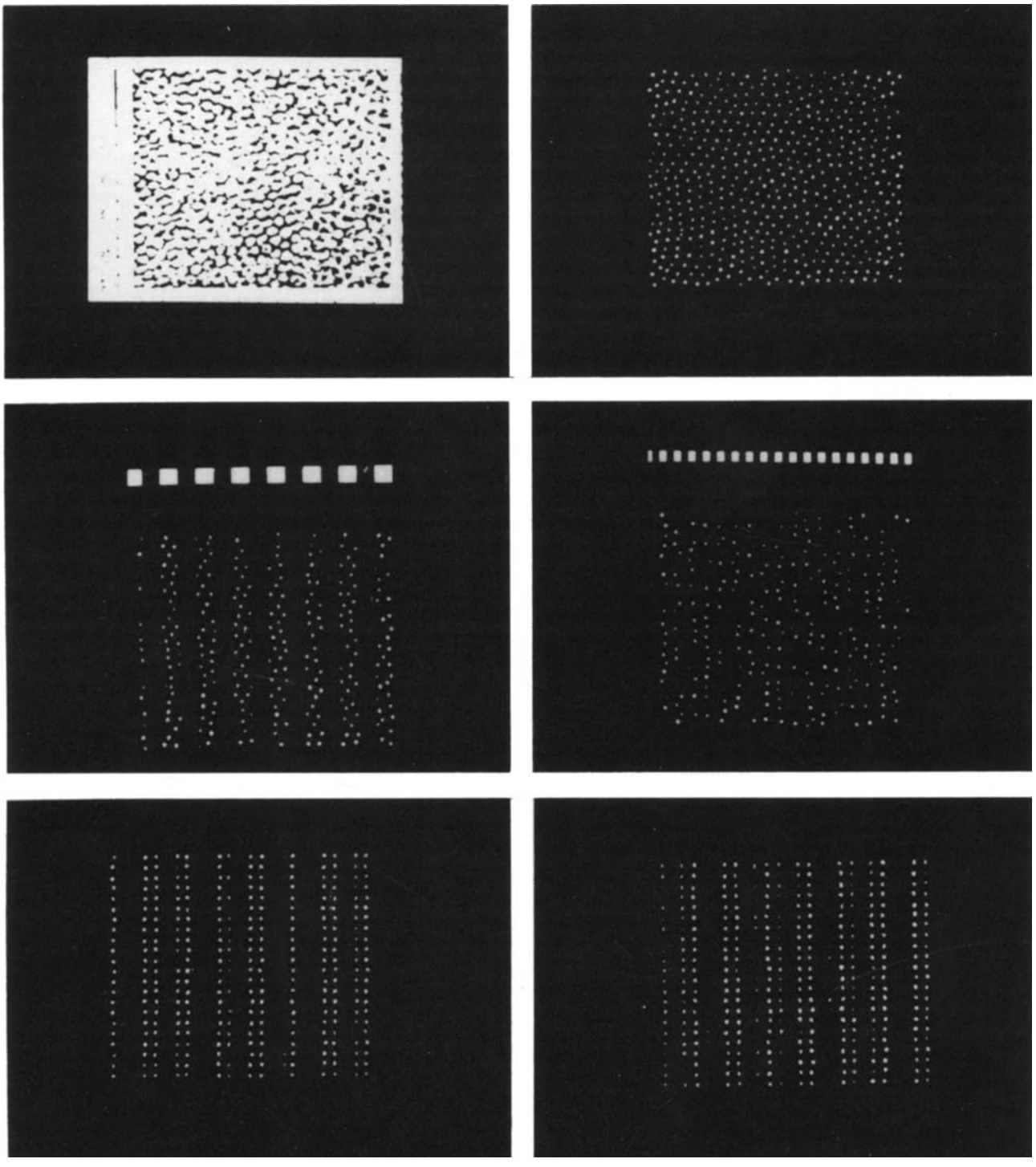

FIG. 18. The *top* and *middle rows* illustrate aliasing by an array of photoreceptors. *Top left*: photomicrograph of a 12′ × 13′ patch of human retina near the center of the fovea. The mean center-to-center distance between receptors is 0.53′, implying an aliasing cutoff of 57 cycles/deg. *Top right*: sampling grid made by punching a pinhole through the center of each receptor in the 12′ × 13′ patch. *Middle left*: 30-cycle/deg square wave grating seen through the pinhole grid. *Middle right*: 80-cycle/deg square wave grating seen through the pinhole grid. Note the appearance of broad curved line segments due to aliasing. The *bottom row* illustrates the effect of a regularly spaced (checkerboard) array of sampling points having roughly the same sample-point distance as the average value for the retinal array. *Bottom left*: 30-cycle/deg square wave grating seen through the regular sampling grid. *Bottom right*: 80-cycle/deg square wave grating seen through the regular sampling grid. [*Top left* from Polyak (98), p. 268.]

CONCLUSIONS. The preceding line of analysis suggests two puzzles. First, because of sampling effects, visual acuity measured by the highest spatial frequency at which a sinusoidal interference pattern can be discriminated from a uniform field should be greater than the local sampling theorem limit everywhere on the retina. Second, spatial "noise" produced by aliased high frequencies should normally be present in the extrafoveal visual field. The fact that neither phenomenon ordinarily occurs perceptually seems to imply that the visual system has some active mechanism for suppressing counterfeit spatial frequencies, i.e., a mechanism that prevents false resolutions that would otherwise contaminate its spatial analysis of the retinal image. As far as we know, current neural models for spatial contrast detection do not explicitly provide such a mechanism, and consequently our ability to trace the flow of spatial information through the visual pathway has a missing link. We will not try to guess the nature of this link, except to make the obvious point that the most natural neural mechanism for suppressing aliased spatial frequencies would be one that relies on information supplied by retinal image motion. It is well known that during normal vision the retinal image is always in motion, and that when motion is artificially stopped vision fails almost immediately. This disappearance effect is usually attributed (in a functional sense) to a kind of neural boredom, but it also seems conceivable that the visual system is not so much concerned with suppressing unchanging spatial signals as with suppressing signals that are spatially ambiguous.

In any event it seems clear that the answer to these questions must come from studies that measure visual acuity under conditions that control retinal image motion. Kelly (75) has recently reported the first measurements of the spatial contrast sensitivity function for gratings that drift across the retina at a controlled velocity independent of the subject's eye movements. His measurements concentrate on lower frequencies and consequently are not directly related to our problem, but they do demonstrate clearly that spatial contrast sensitivity is sharply dependent on the rate of retinal image motion. Visual acuity must now be analyzed from this point of view.

COLOR VISION

Newton's prism experiments of 1666 showed that light comes in different degrees of refrangibility (today we would say different wavelengths or frequencies); that each wavelength in isolation produces its own unique color experience (i.e., the spectral colors); and that lights composed of wavelength mixtures generally produce color experiences indistinguishable from those produced by other mixtures that are physically entirely different (e.g., a red/green wavelength mixture is indistinguishable from a suitably adjusted yellow/blue mixture). For nearly 300 years this last fact was the outstanding problem of color vision. What causes the visual system to lose so much wavelength information? By the middle of the 19th century experiments by Maxwell had established the precise quantitative nature of this information loss, which at daylight levels of illumination takes the form of *trichromacy*: "Given any four lights, whether spectroscopically pure or not, it is always possible to place two of them in one half of a foveal photometric field and two in the other, or else three in one half and one in the other, and by adjusting the intensities of three of the four lights to make the two halves of the field indistinguishable to the eye." [(29), p. 199]. The psychophysical laws of color matching were formalized by the mathematician Grassman in 1853, and it became apparent that the set of color experiences could be modeled as a three-dimensional vector space, in which the three trichromatic "primaries" (that is, three lights of fixed wavelengths which can be weighted in intensity and mixed to match other lights of arbitrary composition) play the role of basis vectors: Just as any point in three-dimensional space can be produced by a linear combination of the vectors (1, 0, 0), (0, 1, 0), and (0, 0, 1), so the color experience generated by any wavelength mixture L can be produced by a linear combination of primary lights A, B, C: Either $L = aA + bB + cC$ (where aA means a quanta/s at wavelength A, and + means the physical superposition of two lights—e.g., by overlapping the beams of two projectors) for some set of positive weights a, b, c, or one primary must be added to L to produce a match, in which case we interpret its weight as negative—so that, for example, $aA + L = bB + cC$. (For a thorough discussion of Grassman's laws and their measurement-theoretic implications, see Krantz, refs. 79, 80.)

Maxwell and Helmholtz (in 89a) independently recognized that trichromacy could be neatly explained by a physiological mechanism for light transduction suggested 50 years earlier by Thomas Young (in 89a). Impressed by the physical impossibility of equipping each retinal point with separate detectors for every wavelength, Young (in 89a) proposed that there might be only three broadband detectors, each corresponding to a "particle," analogous to a resonator, that responded best to some peak wavelength and more or less strongly to other wavelengths according to their distance from that peak. He supposed the peaks of the action spectra of his particles would correspond to the "three principle colors, red, yellow, and blue."

Modern work has exactly confirmed Young's proposal: "His particles are the π-electrons of the chromophores of the visual-pigment units of the cones." [See Rodieck (106), p. 714.] However, for a hundred years after its revival by Helmholtz and Maxwell (see ref. 89a), Young's model could not be physiologically confirmed, and in the absence of any clear understanding of the neural substrate of color vision, competing theories flourished. The most radical alternative (or so it seemed) was Hering's (66a) idea that photorecep-

tors might operate in an opponent-process fashion, with some substance being created by certain wavelengths and destroyed by others. Variations on Young's idea included the addition of more than three cone photopigments; the idea that each type of photopigment is (or is not) segregated into its own private set of receptors; the idea that the spectral sensitivities of the cones might be determined by optics rather than photochemistry, and the notion that the blue cones might be rods.

Many of these hypotheses could be decisively tested by psychophysical experiments (see chapt. 8 in ref. 29), and during the period 1850–1950 color science was largely dominated by psychophysics, which achieved some important successes. In the nineteenth century, perhaps the most notable example was Konig's demonstration (1894; see ref. 77a) that the action spectrum for detection in rod vision agrees with the absorption spectrum of the rod photopigment rhodopsin. That result (since confirmed by modern methods—see Fig. 21) showed that the spectral properties of rod vision at least could be explained by the absorption characteristics of a photopigment. In this century a comparable achievement is the long-term effort of W. S. Stiles to identify the spectral properties of the mechanisms underlying cone vision by the analysis of increment thresholds (120).

From the standpoint of industrial colorimetry, psychophysics alone was sufficient for the practical problem of specifying the color experiences produced by isolated lights of arbitrary spectral composition. [We say isolated because Grassman's laws do not predict the color experiences produced by juxtaposing two lights—e.g., the fact that white light turns pink when surrounded by a green field. These simultaneous contrast phenomena, which Land (82) has demonstrated so effectively, are still not fully understood.] However, psychophysical methods alone could not provide definitive answers to the fundamental questions: Are there really exactly three cone photopigments? If so what are their absorption spectra? Are the photopigment types segregated in different receptors? These questions persisted largely because the cone pigments of the human retina, unlike rhodopsin, could not be successfully isolated (and have not been to this day). The impasse was finally broken in the period 1950–1970 by two new techniques: retinal densitometry, which measures the absorption spectra of visual pigments in the living retina, and microspectrophotometry, which measures the absorption spectra of individual receptors. Both techniques have agreed in showing that besides rhodopsin, the human retina normally contains three cone pigments, which Rushton (112) has called erythrolabe (λ_{max} = 570), chlorolabe (λ_{max} = 540), and cyanolabe (λ_{max} = 440). Microspectrophotometry has shown in addition that these pigments are segregated in different receptors—the (so-called) "red," "green," and "blue" cones. [There are good grounds for objecting to these colorful labels—expressed, for example, by De Valois and De Valois (42) and in the chapter in this *Handbook* by De Valois and Jacobs. However, they are probably too convenient to be readily discarded.]

On the basis of these results—together with converging evidence from a hundred years of psychophysics and physiology—there is today general agreement that the normal human retina contains four photopigments (three cone pigments and rhodopsin), each housed in a separate class of receptors, and that this fact together with the inability of receptors to signal anything beyond their rate of quantum absorption (Rushton's univariance hypothesis) can account for the basic properties of color matching (e.g., see refs. 29, 106, 112, and chapts. 9–15 in ref. 40). The purpose of the present section is to show how the quantitative psychophysical facts of color matching arise from the underlying physiology of the retina. We first consider the special case of rod vision, which serves as an illustration of what can be expected in a monochromatic visual system—i.e., a system in which only one class of receptors is active, all of which share a common absorption spectrum. Then we extend the argument to cover dichromatic and trichromatic systems. We also briefly consider the major forms of color blindness, which can (probably) be explained in terms of losses or spectral distortions of one or more receptor systems. Our discussion does not extend beyond color phenomena that can be explained in terms of the absorption properties of receptors. Consequently we have nothing to say about such important perceptual phenomena as simultaneous color contrast, or about the neural mechanisms that process the color information supplied by the outer segments. On these topics the reader is referred to the chapter by De Valois and Jacobs in this *Handbook*.

Scotopic Spectral Sensitivity

Figure 19 is called the scotopic spectral sensitivity curve of the human eye. It represents psychophysical data, in that it was obtained by systematically varying physical parameters (intensity and wavelength) and measuring the corresponding perceptual effects. To obtain the curve in Figure 19, the subject was first fully dark adapted. Then flashes of light of a given wavelength and adjustable intensity were delivered to his peripheral retina, where rods are most densely packed, and, for each wavelength the intensity was found for which the flash was just on the threshold of visibility (i.e., the intensity for which it was reported to have been seen on 60% of the flashes). In other words, this is a plot of the effectiveness of light on the dark-adapted (scotopic) visual system as a function of wavelength—the scotopic visual action spectrum.

The ocular media absorb some quanta before they arrive at the retina, and the wavelength dependence of this absorption is plotted in Figure 20. When the spectrum in Figure 19 is corrected for the absorption shown in Figure 20, the result is as plotted with

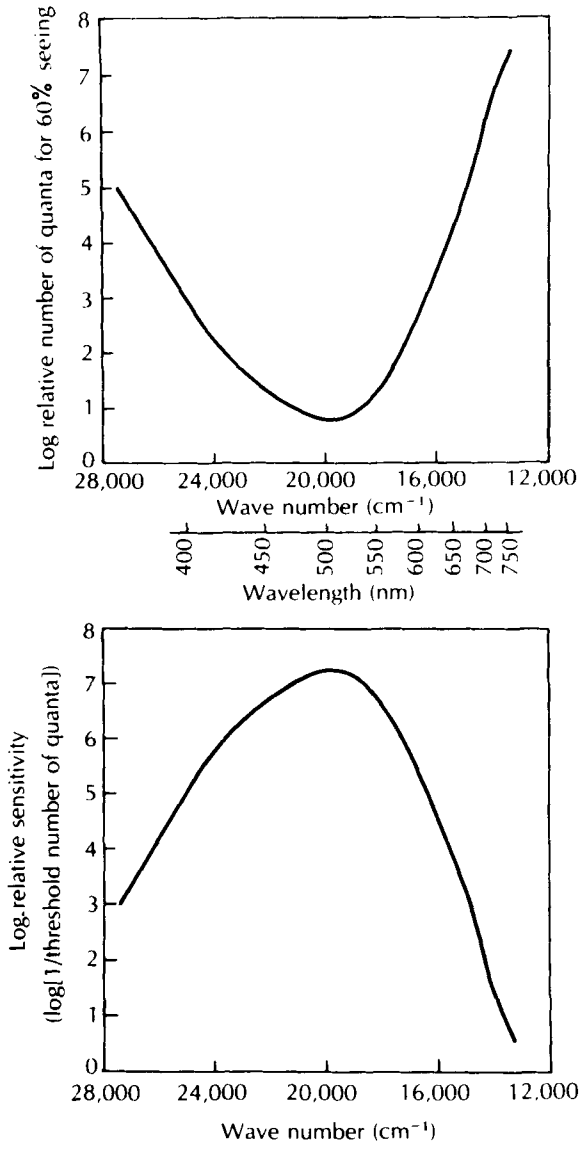

FIG. 19. The scotopic visual action spectrum plotted two ways. Horizontal axis is linear with wave number (number of waves per cm). See text for explanation of data collection procedure. [From Cornsweet (36).]

symbols (×) in Figure 21. It is the action spectrum of that part of the scotopic visual system that is proximal to the ocular media. The dashed curve in Figure 21 is the absorption spectrum of rhodopsin. The coincidence of the dashed curve and the symbols in this figure, that is, the agreement between a psychophysically determined action spectrum and a physically measured absorption spectrum, is strong evidence for a physiological theory, namely, that the absorption of quanta by rods is the initial step in the crucial transduction of light into physiological signals, under scotopic conditions.

Monochromacy

The absorption spectrum of intact rods is primarily determined by the quantal absorption of the rhodopsin molecules within the rods. When a quantum passes through the space occupied by a rhodopsin molecule, it has some probability of being absorbed, and the absorption spectrum is really a plot of this probability as a function of wavelength. Now an extremely important question arises. When a quantum is absorbed by a rhodopsin molecule, does the resulting activity retain information about the wavelength of the absorbed quantum? All available evidence indicates the answer is no. Quanta of different wavelengths (and energies) may have different probabilities of being absorbed, but if a quantum is absorbed, the physiologically significant effect is apparently identical, regardless of the wavelength (or energy) of the quantum. Strong psychophysical evidence that the wavelength information is lost is that, if two patches of differing wavelengths are presented to a dark-adapted eye and their relative intensities are adjusted to compensate for the difference in scotopic sensitivity between the two wavelengths, the two patches will be completely indiscriminable, so long as the intensities of the patches are below the photopic threshold, so that only the scotopic system is operating. Any pair of lights that produce equal numbers of quantal absorptions are indistinguishable from each other. The information that the two patches are different in wavelength must therefore have been lost by the visual system, and that is true for all wavelengths. That information loss evidently occurs during the absorption of the quantum, all absorptions producing identical effects (the *cis-trans* isomerization of rhodopsin) regardless of the energy in the quanta. The differences in energy among absorbed quanta are evidently manifested in small temperature differences that are not signaled to the rest of the visual system.

The assumption that wavelength information is lost at the point of quantum absorption is generally referred to as the principle of "univariance," a term

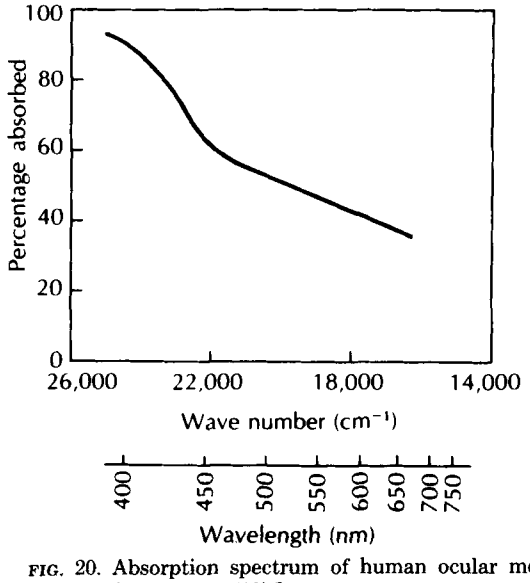

FIG. 20. Absorption spectrum of human ocular media. [From Ludvigh and McCarthy (89).]

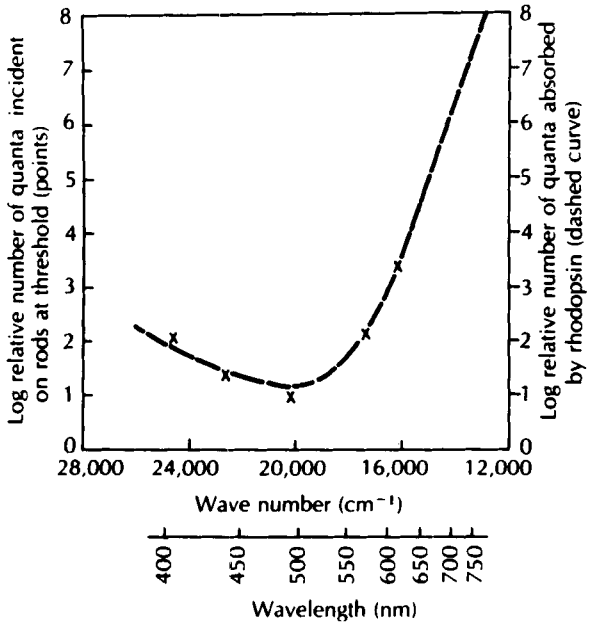

FIG. 21. Scotopic visual action spectrum corrected for absorption by ocular media (×) compared with absorption of rhodopsin (*dashed curve*). [From Cornsweet (36).]

introduced by Rushton (112). Because this hypothesis is central to understanding the relationship between the psychophysics and physiology of color vision, its meaning (and, of course, empirical validity) deserves close examination. Here we take univariance to mean that the effective input to the visual system is the quantum catch of the receptors, so that the response of any given receptor r_0 is some function $f_0(c_0, c_1, c_2, \ldots,)$ whose arguments are the quantum catches of r_0 itself, c_0, and the catches $c_1, c_2, \ldots$ of other receptors $r_1, r_2, \ldots$. Thus two stimuli that produce the same quantum catch in every receptor (i.e., the same values of $c_0, c_1, c_2, \ldots$) constitute identical inputs to the visual system and are necessarily indiscriminable. Note that this is not equivalent to assuming that the response of a receptor depends only on its own quantum catch—the special case in which f_0 depended on c_0 alone, and not on $c_1, c_2, \ldots$. This point is important because there have now been several demonstrations that receptor membrane potentials do depend on the quantum catch of more than one receptor (17, 18, 47) and one sometimes reads that such demonstrations disprove the univariance principle—which clearly they do not.

In fact there seems to be no evidence contradicting univariance. However, this does not mean the principle is undoubtable: Rodieck (106) points out that receptors could provide wavelength information by varying their output signal as a function of the position along the outer segment at which absorption occurs. (For example, quanta absorbed nearer the base might be more effective at modulating transmitter release.) Because the average position at which photons are absorbed depends on their wavelength (the mean distance traveled before absorption is inversely proportional to the absorption probability) a device reading the output to a fixed number of absorptions could make a better-than-chance guess as to the wavelength of the stimulation if it knew where those absorptions had occurred. This would be a violation of univariance because the response of the visual system would depend on something more than quantum catch. However, it does not appear that the human visual system takes any significant advantage of this possibility. (We do not rule out the possibility that the mechanism suggested by Rodieck might produce failures of univariance in a psychophysical experiment designed specifically for the purpose, but we would expect any such failures to be very small.) At the moment then it seems safe to assume that univariance is strictly true and to develop a model of color vision based on that hypothesis. It is unlikely that future developments will force us to alter this assumption in any significant way.

Assuming univariance, lights that produce the same rate of quantum absorption in each receptor must be indistinguishable. To understand what this entails one needs a model of the light-absorption process. The standard first approximation is to assume that the light-catching parts of receptors (e.g., rod outer segments) can be treated simply as tubes filled with photopigment (e.g., rhodopsin) molecules; photopigment concentration is assumed to be uniform throughout the tube, and all quanta are assumed to enter the tube straight-on, and travel down its length until they are either absorbed or reach the end. In this case absorption can be modeled by a spatially homogeneous Poisson process in which the probability density for absorption of a quantum of wavelength λ_i at distance x along the outer segment is $a_i c\, e^{-a_i c x}$, where a_i is the chromophore absorptivity at λ_i (in other words, the intrinsic absorption spectrum of the photopigment), and c is the photopigment concentration. Then if the length of the outer segment is L, the probability, $P(i, c)$, that an incident quantum of wavelength λ_i will be absorbed during its journey down the receptor is

$$P(i, c) = 1 - e^{-a_i c L}$$

$P(i, c)$ is the absorption spectrum of the photoreceptor; it depends on both wavelength and concentration and is nonlinearly related to the intrinsic absorption spectrum of the photopigment.

If I_i quanta/s at wavelength λ_i are incident on the receptor, and the catch rate is small enough to leave the concentration c effectively unchanged, then the rate of absorption will be $P(i, c)I_i$ per second. Assuming that every receptor has the same absorption spectrum (e.g., only rods are active, and all have the same value of the product $a_i cL$), two uniform patches of light of wavelengths λ_1 and λ_2 are indistinguishable if their intensities, I_1, I_2, are adjusted so that

$$P(1, c)\, I_1 = P(2, c)\, I_2$$

Now it is obvious from this equation that at any fixed photopigment concentration level, any pair of monochromatic lights can be made indistinguishable by

adjusting the intensity of one light. However, because the absorption spectrum $P(i, c)$ depends nonlinearly on c, it is also clear that in general, lights that produce equal absorptions at one concentration level will not necessarily do so at other levels. In other words, lights that are indistinguishable at one adaptation level need not match at all adaptation levels. This nonlinearity is formally awkward, and so it is natural to look for simplifying assumptions that can justify getting rid of it. If one requires a model that will guarantee the stability of color matches for all states of adaptation there are only two possibilities. One is that the product, $a_i cL$, is always small enough (≤ 0.2) to justify the approximation $1 - e^{-a_i cL} \approx a_i cL$. In this case two monochromatic lights match at any concentration if their intensity ratio I_1/I_2 equals the ratio a_2/a_1, and if this is true for any concentration it is true for all. Early estimates of the absorption probabilities of individual receptors suggested that $P(i, c)$ could be well approximated by $a_i cL$, but more recent measurements for both rods and cones yield values on the order of 0.5 or more [(106), p. 142–143], so this linear approximation assumption does not seem well founded.

The other alternative is to assume that photopigment concentration changes are negligible over the range of light intensities one is concerned with. In this case $P(i, c)$ can be regarded as a function of wavelength alone and written simply as P_i. Then the condition for indistinguishability of two monochromatic lights, L_1 and L_2, whose wavelengths are respectively λ_i and λ_2 is simply

$$P_1 I_1 = P_2 I_2$$

and for wavelength mixtures (say I_{ij} denotes the number of quanta/s at wavelength λ_j produced by L_i) the matching condition is

$$\sum_j P_j I_{1j} = \sum_j P_j I_{2j}$$

For the rod system this constant concentration assumption seems reasonable, because rods become visually ineffective (i.e., saturate) at intensity levels that bleach only a small fraction (ca. 10%) of the available rhodopsin (1a). For cones, on the other hand, visual matches can be disrupted by very intense adapting lights that bleach substantial fractions of photopigment (29), and so in this case a breakdown of the constant concentration assumption can have significant perceptual consequences. However, throughout most of the intensity range encountered in normal vision such breakdowns do not occur, and because the formal simplification permitted by that assumption is more important in cone vision (due to the additional complications introduced by the three-dimensionality of phototopic vision) it seems reasonable to adopt the hypothesis that for both cones and rods the photoreceptor absorption spectrum, $P(i, c)$, is independent of c, and for each receptor depends only on wavelength. Thus we assume that all receptors of a given class can be characterized by a common absorption probability, P_i, that depends only on the wavelength, λ_i, via the intrinsic absorption spectrum, a_i of the photopigment common to that class. In that case, the requirement for indistinguishability of monochromatic lights by a given receptor system under all conditions is simply $P_1 I_1 = P_2 I_2$, and when several receptor systems are active—as we shall see in the following discussion—the requirement for indistinguishability is a straightforward multidimensional generalization of this equation.

As a consequence of the loss of wavelength information, the scotopic visual system is often said to be totally color-blind. That terminology can be misleading, however. A less ambiguous statement is that the scotopic system cannot discriminate among lights on the basis of their wavelength compositions. If two patches of differing wavelength composition are side by side, it is always possible to adjust the intensity of one until the two patches produce identical effects (numbers of absorptions and isomerizations) on the visual system, and thus to render them indistinguishable. Any visual system for which this is true is said to be monochromatic, or "totally color blind." Note that the term *monochromatic* does not in any sense refer to the perception of colors. If a totally color-blind person asserts that he sees a full spectrum of colors in the world, there is no class of evidence that can refute him. If he is truly monochromatic, what is missing is sensitivity to wavelength differences, and if he were to attribute color names to objects in the world, a manipulation of the intensity and wavelength of light reflected from objects would reveal that there was no *consistent relationship* between the wavelength composition of the reflected light and his color names.

Dichromacy

When the ambient light level is well above the cone threshold, the scotopic system evidently ceases to be functionally significant, and the cone, or photopic, system mediates normal human vision. Suppose that a person had both functional rods and cones, and all of his cones contained a pigment whose absorption spectrum was shifted with respect to rhodopsin, as in Figure 22. At low-light levels, only his rod system would operate and he would be a monochromat. At high levels, only his cone system would be functional, and he would still be a monochromat, since all of the arguments showing that the rod system is monochromatic would apply equally well. [Recent evidence indicates that the rod system can influence perception at relatively high light levels under certain conditions. See, for example, Stabell and Stabell (119). Those conditions are not relevant to the logic of this argument and will be neglected, here, to facilitate the discussion.]

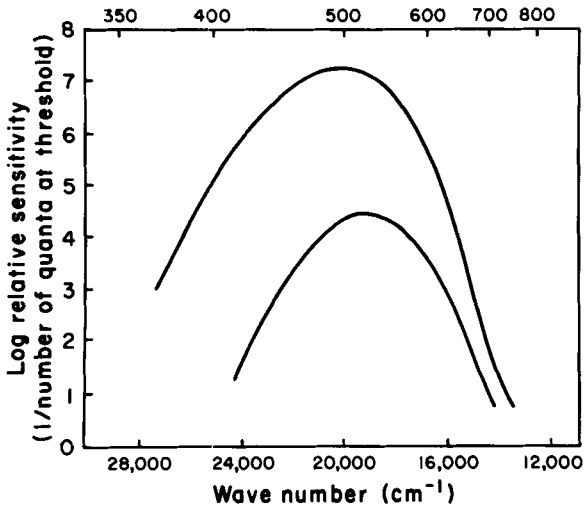

FIG. 22. Spectral sensitivity curve of rods (*upper curve*) compared with spectral sensitivity curve for a hypothetical set of cones (*lower curve*). Note that this lower curve does not represent any real set of cones and is strictly for explanatory purposes.

If a pair of patches of light of differing wavelengths are adjusted to be indiscriminable, that is, to produce equal numbers of absorptions, under scotopic conditions, and then the intensities of both patches are doubled but remain below the photopic threshold, the patches will still have equal effects on the visual system, and so will still be indiscriminable. If the intensities of both patches are now increased by a factor sufficient to raise them well above the cone threshold, the rods will cease contributing to vision and the visual action spectrum will shift to that of the cones. As a consequence, the probabilities of absorption of the various wavelength components will change and the patches will no longer match. However, it will now be possible to readjust the relative intensities of the two patches until they are again indiscriminable, and so the system is again monochromatic.

Now suppose that the intensities of the two patches were lowered until they were above the cone threshold but low enough that the rod system still functioned too. (This intensity range, between pure scotopic and pure photopic vision, is called mesopic vision.) If the intensities of the two patches were adjusted to produce equal absorption in the rods, they would have differing effects on the cones, and so the viewer could tell them apart. If the intensities were readjusted to produce a match for the cones, the rod signal would provide the information necessary to discriminate them. Therefore, a person with rods containing rhodopsin and cones containing a pigment with a different absorption spectrum would be a monochromat at high- or low-light levels, but would not be under mesopic conditions.

A fairly large class of humans (those afflicted with the form of color blindness known as *dichromacy*) have a normal rod system and a cone system that effectively consists of two kinds of receptors, differing in their absorption spectra (see curves for systems A and B in Figure 23*A* and *B*. When these people are presented with photopic stimuli, their visual systems contain two simultaneously operating cone subsystems, and their ability to discriminate wavelength mixtures has the same properties as the person described above under mesopic conditions.

If two patches of differing wavelength composition are adjusted in intensity to have identical effects on the A system, they will have differing effects on the B system, and vice versa. Thus, such a person is not a monochromat under photopic conditions. (He will be under scotopic conditions.) The point labeled 620 in Figure 23*B* is a representation of the effects on the A and B cone systems of a patch of light of wavelength 620 nm and an intensity of 1,000 quanta$\cdot s^{-1} \cdot mm^{-2}$. The location of that point in the two-dimensional space of Figure 23*B* is found simply by multiplying the intensity (1,000) by the proportion of quanta absorbed at 620 nm in each cone subsystem as plotted in Figure 23*A*.

The space in Figure 23*B* is linear. That is, if the intensity of the 620-nm stimulus were doubled, the number of quantal absorptions in each system would double, and a point representing this new stimulus would be twice as far from the origin. Thus, the dashed line in Figure 23*B* represents the effects of all intensities of 620-nm light. As noted above, this linearity only holds exactly as long as the intensity is not so great that a substantial fraction of either pigment is bleached. As a practical matter, this kind of linearity holds for almost all light levels experienced in the world. The space in Figure 23*B* represents the joint consequences of the action spectra of two visual subsystems, and we will call it the visual action space.

Suppose a second stimulus is now placed next to the first, at a wavelength of 560 nm and an intensity of 1,000 quanta$\cdot s^{-1} \cdot mm^{-2}$. Its effect on the A and B systems is represented by the dot labeled "560" in Figure 23*B*. The fact that this dot is not coincident with the point labeled "620" (i.e., that the effects of this stimulus are different from those of the first one) means that the two stimuli are discriminable (provided that the differences between their effects are not lost at any later stage in the system). Furthermore, changing the intensity of either patch would simply move its effect along the line connecting it with the origin, and thus there is no intensity adjustment that will render the two stimuli indiscriminable.

Now suppose that the patches are left the same, 1,000 quanta$\cdot s^{-1} \cdot mm^{-2}$ at 560 nm and 1,000 quanta$\cdot s^{-1} \cdot mm^{-2}$ at 620 nm, but a new component, 1,000 quanta$\cdot s^{-1} \cdot mm^{-2}$ at 500 nm, is added to the 620-nm patch. The effect on the visual system of this mixture (620 nm + 500 nm) of lights can be represented simply as the vector sum of the effects of each of its components, as in Figure 24, because the action space is

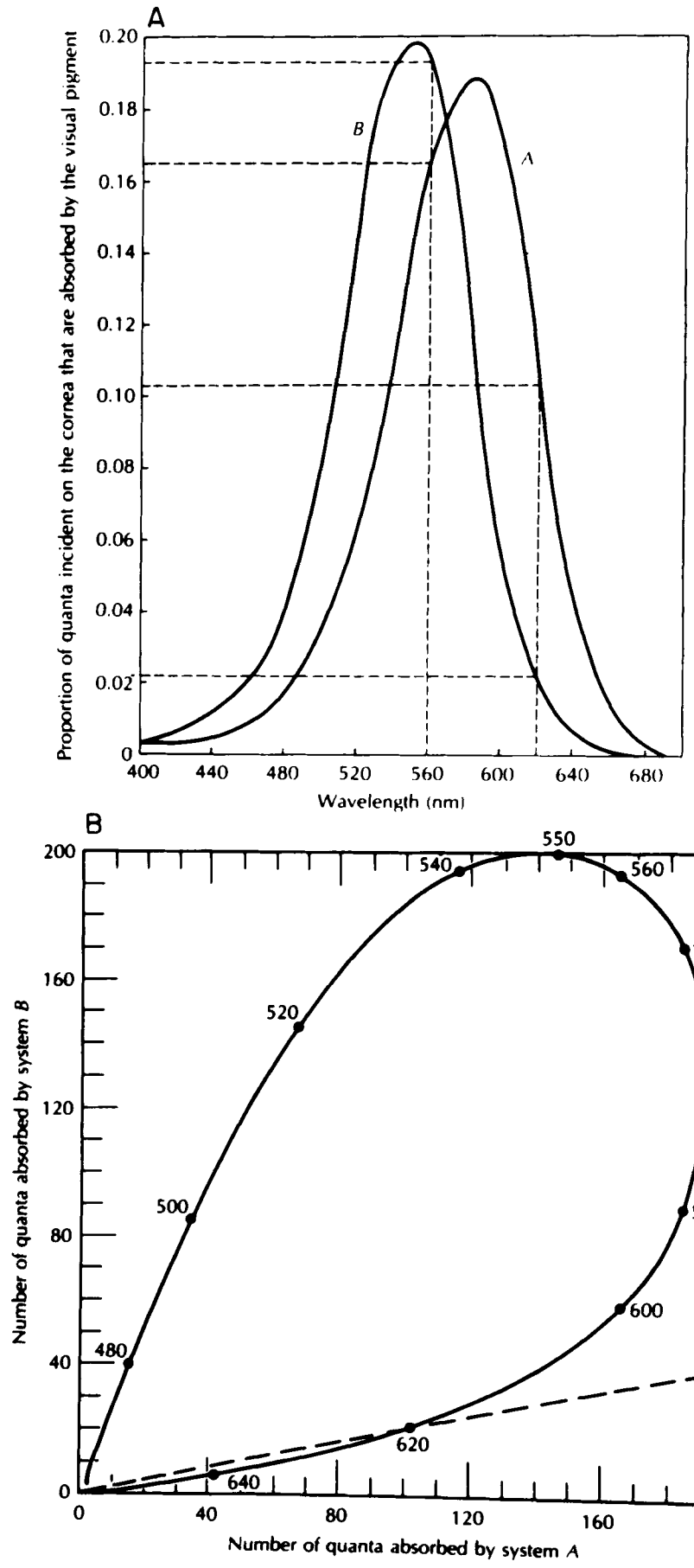

FIG. 23. *A*: absorption spectra for two classes of cones. (These are two of the three classes present in the normal human eye.) *B*: two-dimensional visual action space of a person whose eyes contain only the two kinds of cones plotted in *A*. See text for derivation of system B from A. [From Cornsweet (36).]

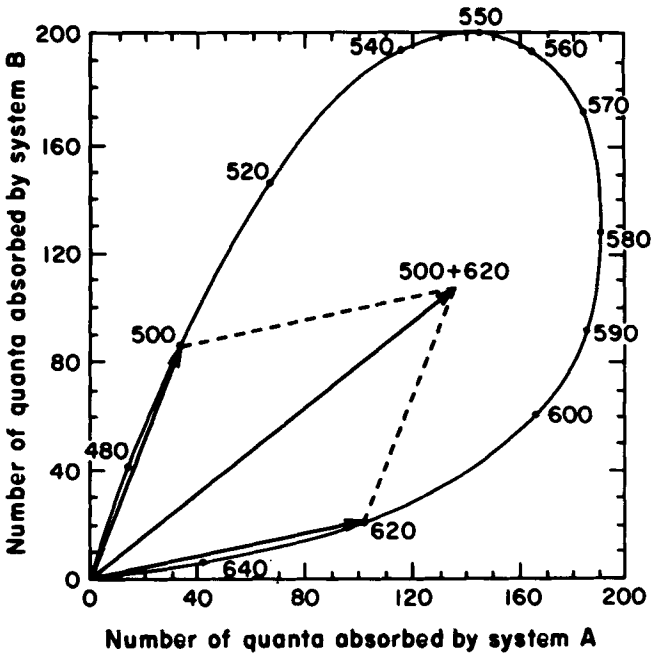

FIG. 24. A representation in two-dimensional action space of the effects of adding lights together. The result can be represented simply as the vector sum of the effects of the components.

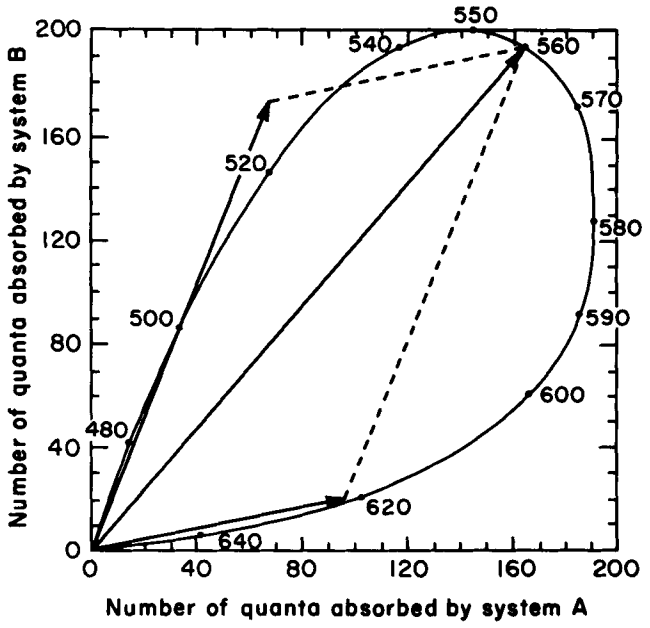

FIG. 25. The effects of a mixture of lights of wavelengths 500 nm and 620 nm and of appropriately chosen intensities (2,050 and 910 quanta·s^{-1}·mm^{-2}, respectively, in this example) will be identical to the effects of a light of 560 nm (at an intensity of 1,000 quanta·s^{-1}·mm^{-2}).

linear—that is, because each wavelength component produces a number of absorptions that is independent of the absorptions produced by the other component. The total number of absorptions in each pigment system is simply the sum of the absorptions that would have been produced by each wavelength component alone.

Now suppose that one patch delivers 1,000 quanta·s^{-1}·mm^{-2} at 560 nm and the other is a mixture of 2,050 quanta·s^{-1}·mm^{-2} at 500 nm and 910 quanta·s^{-1}·mm^{-2} at 620 nm. Figure 25 shows that the vector sum of the 500 and 620 nm mixture is represented by the same point in action space as is the patch of 560 nm at 1,000 quanta·s^{-1}·mm^{-2}. Because the effects of the mixture on both the A and B systems are identical with the effects of the 560-nm stimulus, the two patches must be indistinguishable. Furthermore, it is easy to see that, given a mixture of any two wavelengths, one shorter and one longer than 560 nm, intensities can be found for the two components of the mixture such that their vector sum equals the vector for the 560-nm patch. If it were possible to present stimuli of negative intensity (the physical representation of which will be discussed later), it should be clear that this property would not be restricted to a mixture of wavelengths one on each side of 560 nm, nor restricted to a mixture of just two wavelengths. In general, if negative intensities were possible and if a visual system contained just two classes of receptors whose action spectra were different, so that the effects of any mixture of wavelengths could be represented in a two-dimensional action space like that in Figure 23*B*, then the effects of any mixture of wavelengths could be exactly matched by the effects of any other mixture of wavelengths if the intensities of two different wavelength components were properly adjusted. The effects of each mixture are found by vector addition of its components. The effects of the two mixtures are represented by two points in action space. By changing the intensity of any one wavelength component of a mixture, the effect of the mixture can be moved along a line parallel with the line through the origin representing the effects of that particular wavelength. If the intensity of a different wavelength is changed, the effects of the mixture move in a different direction. Therefore, by changing the intensities of any two wavelengths, the effects of a mixture can be placed anywhere in the two-dimensional action space; that is, any mixture can be made to match any other mixture by the adjustment of the intensities of any two components.

Consider the simple case illustrated in Figure 26, for which a match is to be made between a 620-nm patch and a mixture of 540 nm and 580 nm. If the match is to be made by adjusting the intensities of the 540-nm and 580-nm components, the 580-nm component must have an intensity of 790 quanta·s^{-1}·mm^{-2}, but the 540-nm component must have an intensity of −425 quanta·s^{-1}·mm^{-2}, which is not physically realizable. However, if the 540-nm component were moved from the patch containing the 580-nm component to the 620-nm patch and given an intensity of +425, the two patches would match. In general, if a match between two patches requires a negative intensity as represented in the action space, the match can be physically

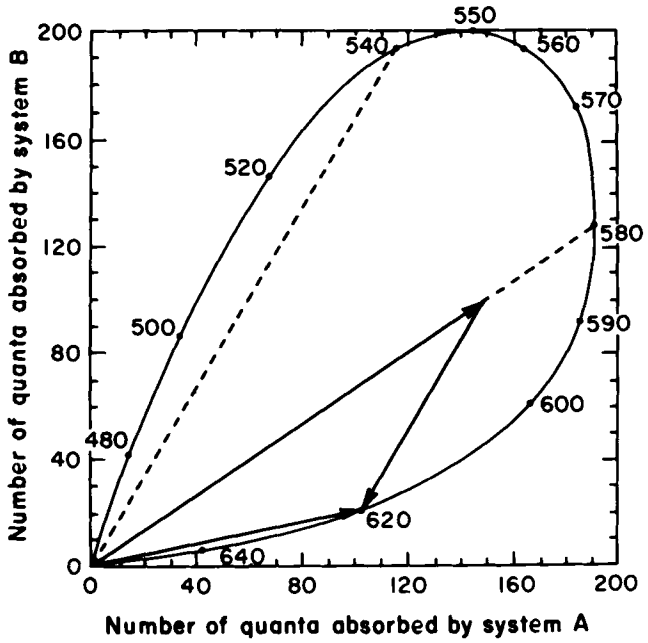

FIG. 26. If a mixture of two wavelengths is to match a third whose effects do not lie on a line between the two to be mixed, one of the two components of the mixture must have a negative intensity, as shown by the downward direction of the 500-nm vector. The physical realization of this negative vector requires the addition of 500 nm light of that magnitude to the 620-nm component.

achieved by changing the sign of the intensity and adding it to the other patch instead.

The action space in Figure 23*B* is two dimensional because it represents the actions of lights on two systems, each of which is one dimensional (that is, the effect of light on each system depends only on the total number of quantal absorptions, not on their wavelengths). It follows that any visual system containing only two classes of receptors that differ in their absorption spectra must have the following property: any mixture of wavelengths can be exactly matched by any other mixture if the intensities of any two component wavelengths can be arbitrarily adjusted (including the realizable equivalent of negative intensities). A visual system with those properties is called dichromatic.

Trichromacy

Most humans cannot produce a match between two mixtures by adjusting the intensities of only two components. Three adjustments are required. That is, normal human vision is trichromatic and it must be represented by a three-dimensional action space. This is to be expected from the fact that the normal human visual system contains receptors with three different action spectra. One estimate of these spectra is shown in Figure 27, and Figure 28 is a representation of the corresponding three-dimensional action space.

If a normal human observer is shown two patches of light, one containing any arbitrary mixture of wavelengths and the other any three wavelengths (except for a special case to be described below), and if the intensity of each of the three can be adjusted (including negative values, i.e., added as positive quantities to the other patch), it will always be possible to find a set of three intensities such that the two patches are indistinguishable. In three-dimensional action space, this is represented in the following way. The three actions of light (on the three color systems) of any single wavelength component are represented by a vector through the origin whose length is proportional to intensity and whose direction depends upon wavelength, and the actions of any mixture of wavelengths can be represented by the vector sum of its component vectors. Thus a patch containing an arbitrary mixture is represented by some vector or point in the action space. Given any set of three wavelengths in a second patch, the sum of their vectors can always be made to equal the vector representing the arbitrary mixture, if each intensity can be adjusted. However, this statement is only true so long as the vectors representing the three wavelengths do not all lie in the same plane. If they did, then a mixture of any two of them could be made indistinguishable from the third by the appropriate intensity adjustment. In other words, for

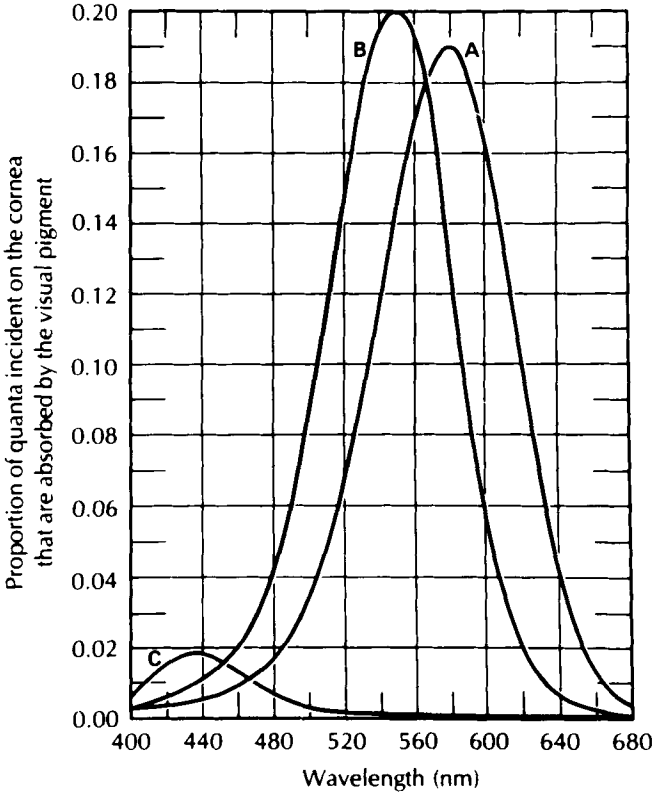

FIG. 27. Estimates of the absorption spectra of the three normal human cone pigments. [From Wald (131). Copyright 1964 by the American Association for the Advancement of Science.]

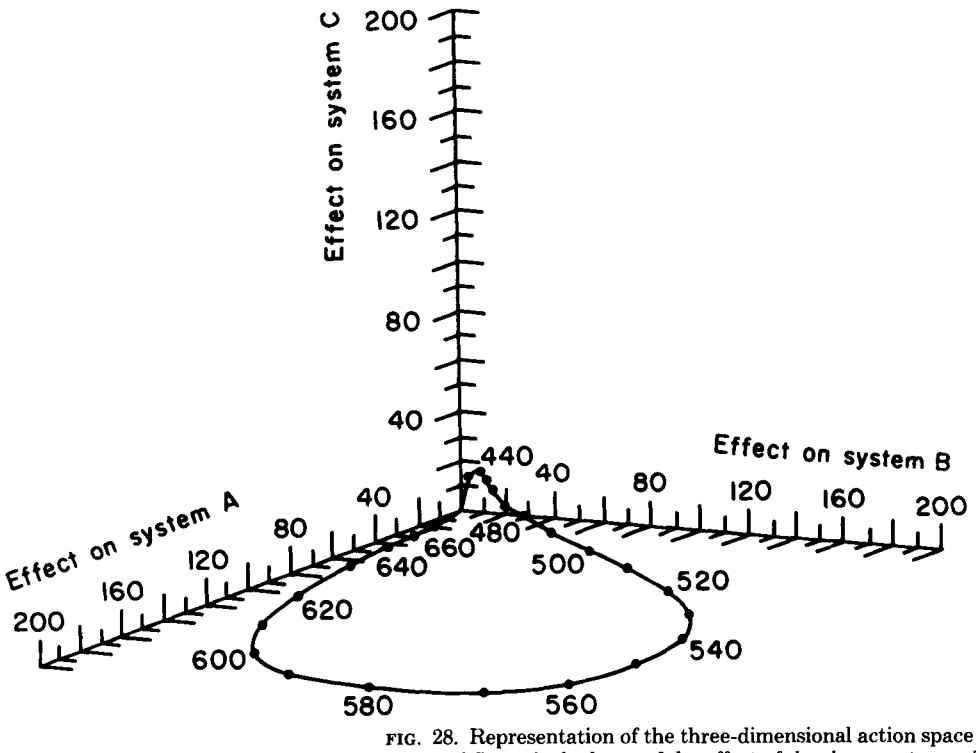

FIG. 28. Representation of the three-dimensional action space of the normal human visual system. The curved figure is the locus of the effect of the three systems of all wavelengths at a fixed intensity. [From Cornsweet (36).]

three wavelengths whose vectors all lie in the same plane, the three-dimensional visual system is, in effect, two dimensional. That condition can occur in normal human vision only in the long-wavelength region of the spectrum, where the blue-sensitive system provides a negligible contribution.

Color Blindness

The three-dimensional space in Figure 28 represents the actions of lights on the normal or trichromatic visual system. If the system were lacking any of the three normal color systems, its action space would collapse to a two-dimensional one, with axes corresponding to the two remaining color systems. This is precisely what seems to happen psychophysically in the type of "color blindness" known as dichromacy. By definition, a dichromat is a person who can match every light (i.e., any wavelength mixture) by adjusting the intensities of only two primaries, rather than the normal three. This color deficiency is further subdivided into three categories: protanopes (who need only a green and a blue primary), deuteranopes (who need only red and blue), and tritanopes (who need only red and green). Protanopia occurs in about 1% of males, deuteranopia also in about 1%, and tritanopia in 0.002%. (However, as noted earlier, everyone is apparently tritanopic in a tiny region at the center of the fovea.) For females (144) the corresponding frequencies are much lower for protanopia (0.02%) and deuteranopia (0.01%), but about the same for tritanopia (0.001%).

The simplest physiological explanation for this psychophysically defined color deficit is that one of the three cone types is missing: protanopes presumably lack the long-wavelength sensitive cones, deuteranopes the middle-wavelength sensitive cones, and tritanopes the short-wavelength sensitive cones. This idea is called the receptor-loss hypothesis and dates back to a suggestion in 1807 by Thomas Young (145). For many years its major competitor was the hypothesis that two of the normal cone systems might be neurally fused into one at a very early stage, thereby effectively creating a single system whose action spectrum would be a weighted average of its two components. Retinal densitometry seems to have disproved the latter hypothesis, since it shows that protanopes lack a long-wavelength photopigment, and deuteranopes lack a middle-wavelength pigment (112). Consequently, today the receptor-loss hypothesis is the generally accepted explanation of dichromacy.

However, recent work indicates that the classic version of that hypothesis needs to be amended somewhat. Suppose one assumes that not only do dichromats lack one of the three cone systems, but in addition their two remaining systems have the same action spectra as those of normal trichromatic observers—who in turn all share identical action spectra for all three of their cone systems. If we perform a confrontation experiment in which a trichromat adjusts the

intensities of three primaries to produce a mixture M that matches any given light L, this expanded version of the receptor-loss hypothesis predicts that any dichromat must also perceive a match between M and L. This is so because the normal observer's match implies that M and L produce the same quantum catch in all three cone systems, and consequently they must also produce the same catch in whichever two of those systems are possessed by the dichromat.

By the same reasoning, the converse experiment should turn out differently: two patches that match to a dichromat will generally not match to a trichromat. Consider the set of mixtures represented by any line parallel to the axis representing system A in Figure 28. These mixtures have differing effects on a trichromatic visual system and can therefore be discriminated. However, because their effects lie along a line parallel to the axis representing the long-wavelength sensitive system (that is, they differ only in their effects on the long-wavelength sensitive system) a protanope would not be able to distinguish among them. (This line is called a *protanopic color-confusion line.*) Similarly, all mixtures whose effects would be plotted along any line parallel to the axis representing the middle-wavelength sensitive system would be indiscriminable to a deuteranope, etc. Thus dichromats cannot distinguish between the colors of certain objects that look different to most people, and so they are called *color-blind.*

Curiously, the idea that dichromats accept the same color matches as normal trichromats was for many years generally accepted, and widely cited as a proof of the receptor-loss hypothesis, even though there was really not much evidence to support it. Recently, Alpern and his collaborators (3–5) have made a systematic study of the question and have shown that in general, dichromats do not accept trichromats' matches, neither do they generally accept the matches of other dichromats. The differences between observers are fairly small, but well outside the bounds of measurement error.

To understand dichromacy in light of these results, one needs to consider the other major form of color blindness: the so-called trichromatic anomalous defects, which occur in about 5.5% of the male population and 0.4% of females (144). Anomalous trichromats are people who require three primaries to match every light (and consequently are trichromatic by definition), but whose intensity adjustments of those primaries are significantly different from those of "normal" trichromatic observers. Like the dichromats, anomalous trichromats are subdivided into three classes: protanomalous, deuteranomalous, and tritanomalous, according to the spectral range in which their differences from the norm are most apparent. If a patch of light containing an arbitrary wavelength mixture is to be matched by a mixture of short-, middle-, and long-wavelength primaries, a protanomalous observer requires a more intense long-wavelength primary than does a normal observer; a deuteranomalous requires a more intense middle-wavelength primary, and so on. It is as if the protanomalous observer has a long-wavelength system whose action spectrum is shifted down toward lower wavelengths, so as to be relatively less sensitive to red light.

Now classically the anomalous trichromats were sharply distinguished from both dichromats and normal trichromats, as though they corresponded to three completely different underlying conditions: trichromats all had three "normal" cone systems; dichromats had only two of the three normal systems, and a third whose action spectrum was shifted away from the norm. The results of Alpern et al. (3–5) suggest a somewhat different story: Apparently there is random variation across the population in the exact location of the action spectra of each color system—as though every person constructed his own idiosyncratic visual system by sampling randomly from a collection of possible long-wavelength action spectra, and again from a collection of possible middle-wavelength spectra, and so on. In the continuum model, a dichromat is simply an unlucky person who happened to select, for example, a long-wavelength system that has effectively the same action spectrum as his middle-wavelength system, whereas an anomalous trichromat is not quite so unfortunate, having selected, for example, a long-wavelength action spectrum that is substantially shifted away from the model position but still does not entirely overlap the spectrum of his middle-wavelength system. "Normal trichromats" are those to whom chance has allocated three systems with action spectra that are all close to the modes of their respective populations. This model thus retains the essential feature of the receptor-loss theory of dichromacy—the idea that the dichromat has only two distinct cone action spectra instead of three—but blurs the old sharp distinctions between classes of normal and color-deficient observers, replacing them instead with purely statistical categories.

Dichromats are called *color-blind* not because they don't see colors, but because they confuse colors that appear clearly different to people with "normal" color vision. However, from an absolute standpoint, "normal" color vision is only a little less color blind than dichromacy. There is still an infinite set of mixtures of wavelengths that physically differ from each other and yet are indistinguishable to the trichromat. A monochromatic visual system loses all information about the wavelength composition of a stimulus, a dichromatic system loses a lot but not all such information, and the normal trichromatic system loses a little less than the dichromatic system.

The very large loss of wavelength information that occurs in trichromatic vision is exploited in many aspects of technology, especially in the printing, photographic, and television industries. That very long story is briefly mentioned here. If a television technician wants to show the audience a scene containing, say, a banana, and if the scene is to be viewed by an

imaginary person whose visual system has lost no wavelength information (e.g., if his retina contains receptors with an infinite variety of action spectra and the rest of his nervous system retains information related to the differences among the outputs of these systems), the technician will have to deliver to the observer's eye a mixture of wavelengths that is identical to the mixture reflected by a banana, and that is, in fact, a very complex and continuous spectrum spanning the entire visible range of wavelengths. Thus his camera will have to be able to detect every wavelength independently, the screen of the receiver will have to contain an independently excitable phosphor that emits at each visible wavelength, and the electronics will have to permit the appropriate independent excitation of each of those phosphors. However, if the observer is as color-blind as a trichromat, the camera, receiver and electronics can be much simpler. Thus the typical studio color TV camera has three camera tubes within it, each viewing the same scene through a differently colored filter, so that there are three sensing systems whose action spectra are different. Similarly, the color receiver screen contains only three different phosphors each emitting light in a different region of the spectrum, and the electronics permit the relatively independent control of intensity of excitation of each.

Note that, because there is no operation in television reproduction that corresponds to the production of negative intensities, once any given set of phosphors has been chosen, there are some colors that cannot be reproduced. Referring to the action space in Figure 28, if the three phosphors each were to emit monochromatic light of a different wavelength, then the receiver can exactly reproduce any color whose action can be plotted as a point somewhere within the volume included in a triangular pyramid whose vertex is the origin and whose edges are the lines representing the effect of each of the three emitted wavelengths. However, the volume that includes all possible colors is roughly cone-shaped and convex everywhere except in the extremely long wavelength region of the spectrum; it includes and is larger than the pyramid-shaped volume that is enclosed by any finite set of wavelengths. In the television industry, phosphors are chosen to emit wavelengths whose vectors in action space include as much of the entire spectral volume as possible.

Seeing Colors

The reader will have noticed that in most of the preceding discussion we have refrained, sometimes with evident awkwardness, from mentioning any color names, like red, blue, and so on. In fact, none of the evidence or arguments presented so far has dealt with questions about the actual perception of colors. We have been concerned exclusively with discrimination or lack of it among patches of differing wavelength composition, and, paradoxically, the only definitions of the various forms of color blindness that are unambiguous are those that refer to discriminations, not to colors seen. However, once the preceding foundations have been established, many aspects of the perception of colors become easier to describe, systematize, and understand.

Each point in the trichromat's action space is a representation of the effects of some light, or mixture of lights, on his three color systems. Imagine the following simple procedure. We present various mixtures of wavelengths and intensities to a trichromat, mixtures represented by a broad sampling of points in the action space, and simply ask him to describe each mixture. (Each mixture is presented as a steadily illuminated patch on a dark background.) We can then map his descriptions onto the volume of action space and look for correspondences between his perceptions and the various combinations of actions on the three color systems. When this is done, the map displays certain regularities. The most obvious ones are: *1*) as the effects of a mixture move away from the origin, the stimulus is called "brighter," and *2*) all stimuli plotted along any straight line running through the origin are given the same hue name, e.g., orange, and stimuli not on the same line through the origin are given different names. Evidently, the hue name is somehow closely correlated with the *ratios* of excitations of the three systems and brightness is correlated with the strength of excitation. (Note that these statements are only approximately correct. A line that is the locus of all mixtures given the same hue name is generally not quite a straight line, and points equally far from the origin but at different places in action space are not necessarily called equally bright.)

From a phenomenological point of view, neither of these properties of the map is surprising. If an object is illuminated with a light of varying intensity, all wavelength components of the reflected light vary in proportion to the intensity of illumination and so the ratios among the intensities of the reflected wavelength components are constant, and the map simply says that, under these conditions, the hue of the object will stay the same and its brightness will vary. But when the statement is seen in the context of action space, one is led to wonder about the physiological mechanism that yields a constant response, e.g., "yellow," when all the inputs are varying but maintaining a constant set of ratios. No physiological mechanism that performs such an operation is obvious. [One plausible mechanism performs a transformation on each of the three quantum catches that is approximately logarithmic and then performs subtractions between the transformed outputs (14, 49).]

Now consider a line running through the origin and forming equal angles with all three axes in Figure 28, and then imagine a section through the trichromatic action space that is a plane perpendicular to that line. Such a plane is represented in Figure 29. The trian-

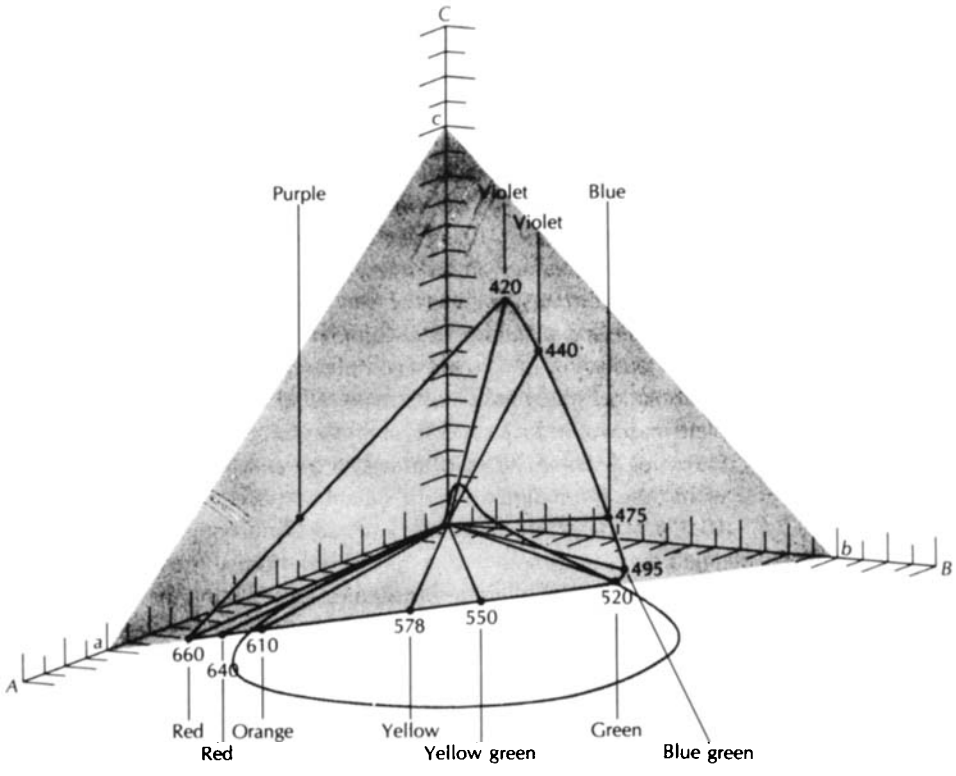

FIG. 29. A plane through trichromatic action space, showing the hue names of some of the stimuli it represents. The *slightly curved triangular line* in the *shaded plane* is the locus of the intersections of lights of all wavelengths with the plane. Therefore, it encloses a space representing the effects of all possible mixtures of wavelengths. [From Cornsweet (36).]

gularly shaped line in this plane can be called the spectral locus. It is the locus of the intersections in action space between this plane and the lines representing every wavelength in the visible spectrum. Stimuli whose actions can be plotted as points in this plane must lie within the spectral locus, and they will be called by various hue names, some of which are listed in Figure 29. To the extent that the lines of constant hue in action space are straight and run through the origin, all sections of action space that are parallel to this one can be represented by maps that differ only in magnification or scale factor, and therefore, to the extent that lines of constant hue are straight, the two-dimensional map in Figure 29 completely represents the relationships between the actions of lights on the color systems and the resulting hue names. In other words, because there are three subsystems, action space is three dimensional, but because hue names are correlated well with ratios of color system excitation, the space representing hue is only two dimensional, the dimension representing intensity being collapsed. (Similarly, to the extent that brightness is correlated with distance from the origin, the space representing brightness is one dimensional.)

If the angle of the plane in Figure 29 were changed, e.g., if the plane were made parallel to one containing two of the axes, the shape of the hue map would change correspondingly, but all the arguments above would still apply. Furthermore, if the action space were plotted so that the three axes were not mutually perpendicular, the shape of the hue map would change, but its important properties would not. The familiar C.I.E. (Commission Internationale de l'Éclairage) diagram, extensively used to specify hues, is one of these linear transformations of a section through action space.

Now suppose that two lights of wavelengths 640 nm and 490 nm are mixed together. As the ratio of the intensities of the two components is varied, the vector in action space that represents their mixture will move in a plane defined and bounded by the 640-nm and 490-nm vectors as shown in Figure 30. To the extent that lines of constant hue are straight in action space, the hue names for all of the mixtures of 640 nm and 490 nm can be represented along a straight line in the two-dimensional hue map, the line of intersection between the plane defined by the 640-nm and 490-nm components and the arbitrarily chosen hue plane. This line, and the hue names associated with various points along it, are shown in Figure 31. If two different wavelengths are now mixed, for example 440 nm and 578 nm, there will be some ratio of their intensities for which the mixture is represented by a vector whose angle is coincident with a 640-nm + 490-nm mixture, and it is seen as the same hue, as represented by the point in Figure 31 where the two lines cross.

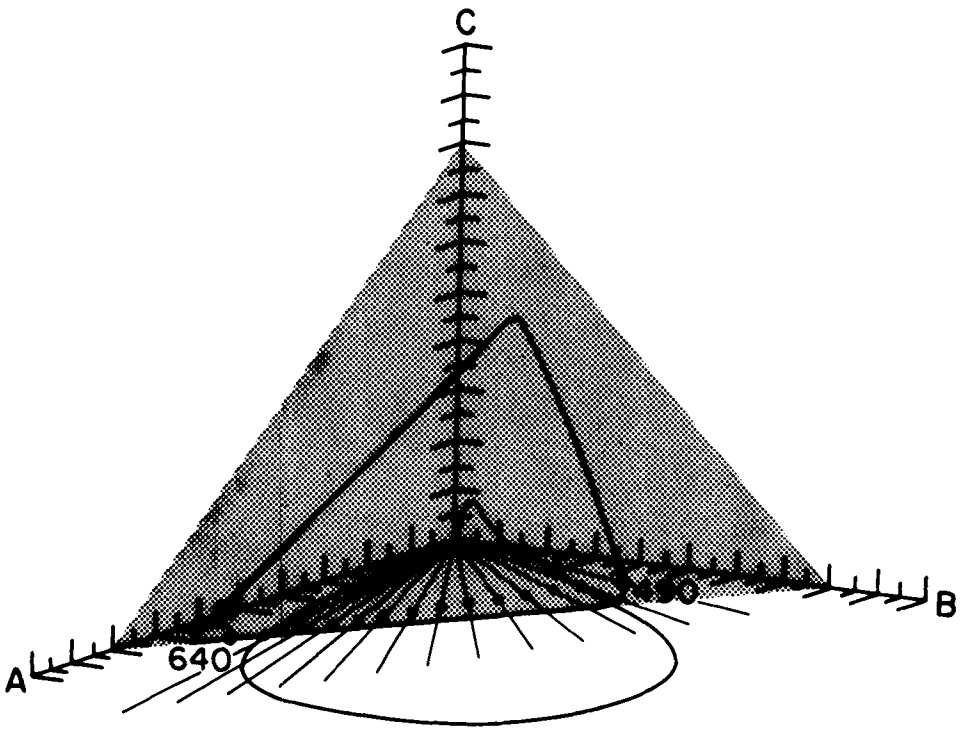

FIG. 30. A representation of the effects of mixtures of lights at 640 nm and 490 nm. As the relative intensities of the two components change, the effect of the mixtures move in a plane indicated by the *straight lines* emerging from the origin. The intersections of these lines with the *shaded plane* are indicated by *dots*, and form a straight line between the points representing 640 nm and 490 nm.

VISUAL ADAPTATION

Overview

As we pass from sunlight to shade or as we emerge from darkness to day, vision is at first difficult. But with time, our eyes adjust to the new surroundings. The adjustment of the visual system to changes in illumination is called visual adaptation. In this section we discuss the nature of this adjustment, this adaptation.

Three problems have dominated the psychophysical analysis of visual adaptation. The first problem—anatomical in nature—has been to identify the class of photoreceptors that mediates detection of lights at threshold. The second—also anatomical—has been to identify the class of photoreceptors that controls the observer's sensitivity to threshold lights. These two problems are distinct, for even though an observer may detect a signal originating in a single class of photoreceptors, sensitivity to that signal may be influenced by the adapted state of other photoreceptor classes via lateral and proximal interactions. The third problem is to model the physical and physiological mechanisms of visual adaptation. The analysis of the mechanism of adaptation stands as a central open problem in visual science. The history of the analysis is rich, and we devote more than half of our discussion to this problem.

A complete review of the research devoted to these three problems requires far more space than we can allot. We limit our discussion to research conducted under scotopic viewing conditions, that is, conditions where detection depends upon a rod-initiated signal.

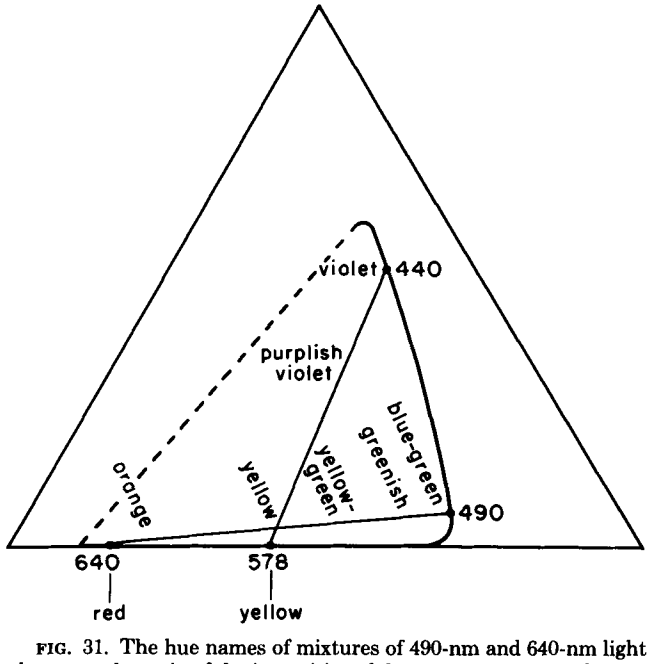

FIG. 31. The hue names of mixtures of 490-nm and 640-nm light change as the ratio of the intensities of the two components change, as indicated along the *line* between 640 nm and 490 nm. There is a particular mixture of 440-nm and 578-nm lights that will produce an identical hue, as indicated where the two straight lines cross.

The organization of this section is as follows. In the next section we briefly describe the basic, experimental paradigms of psychophysical adaptation, i.e., the increment-threshold and dark-adaptation experiments. In *Identifying Neural Substrates,* this page, we review the major findings concerning the identification of the neural substrates that influence detection of weak test lights.

In *Single-Variable Theories,* p. 296, we turn to a discussion of the mechanisms of adaptation. In particular, we review the status of the hypothesis that the state of adaptation in a small region of the retina may be characterized by a single real number. Many classic theories of adaptation (e.g., refs. 9, 64, 111, 121, 130) have incorporated this assumption. Its rejection, which is likely though not yet certain, is largely responsible for the current interest in multiple channel theories of adaptation (see ref. 90).

Experimental Paradigms

Consider the following experiment: An observer centers his gaze on a large (say, 10°), uniformly illuminated field of light, called the conditioning field. The observer's threshold to a small, brief, test flash is measured. The conditioning field may be presented steadily, and the observer allowed to fully adapt, before threshold is measured. This is the equilibrium case. The experiment is then called a *light-adaptation* or *increment-threshold* experiment.

The classic increment-threshold data collected by Aguilar and Stiles (1a) under scotopic viewing conditions is shown in Figure 32. As can be seen in the figure, threshold rises as the intensity of the conditioning field rises. Threshold intensity is proportional to conditioning-field intensity over a sizable range. This part of the increment-threshold curve is generally referred to as the Weber's law region.

When the conditioning field is made to vary over time, the observer's adapted state also varies. Threshold may then be measured relative to the changing conditioning field. This is the dynamic case. A most important dynamic experiment is the measurement of threshold following the offset of a previously steady conditioning field. The experiment is then called a *dark-adaptation* or *recovery* experiment. Examples of dark-adaptation curves are shown in Figure 33.

In both the light- and dark-adaptation experiments, the conditioning field establishes the ambient illumination at the observer's eye, i.e., the action of the conditioning field—its adapting effect—is the object of study. The test light is used to probe the effect of the conditioning light upon the observer's eye; i.e., the test serves as a measuring device, and it is assumed to be too weak to disturb the state of adaptation established by the conditioning field.

Identifying Neural Substrates

Earlier in *Overview,* p. 293, we outlined three basic problems that must be solved before an adequate understanding of scotopic adaptation may be claimed. First, the psychophysical parameters that cause detection to be mediated by a rod-initiated signal must be specified. Knowledge of these parameters can enable us to assert that the scotopic pathway has been well and truly isolated. Second, the class, or classes, of photoreceptor signals that govern the state of adaptation of the scotopic pathway must be identified, because, although detection may be initiated by rods alone, adaptation may depend on both rods and cones. Finally, the mechanisms of adaptation that determine the pathway's sensitivity must be described. The strongest proposal in this regard is the hypothesis that different states of adaptation may be described by the value of a single variable.

It is important to realize that the resolution of the last problem is logically independent of the other two.

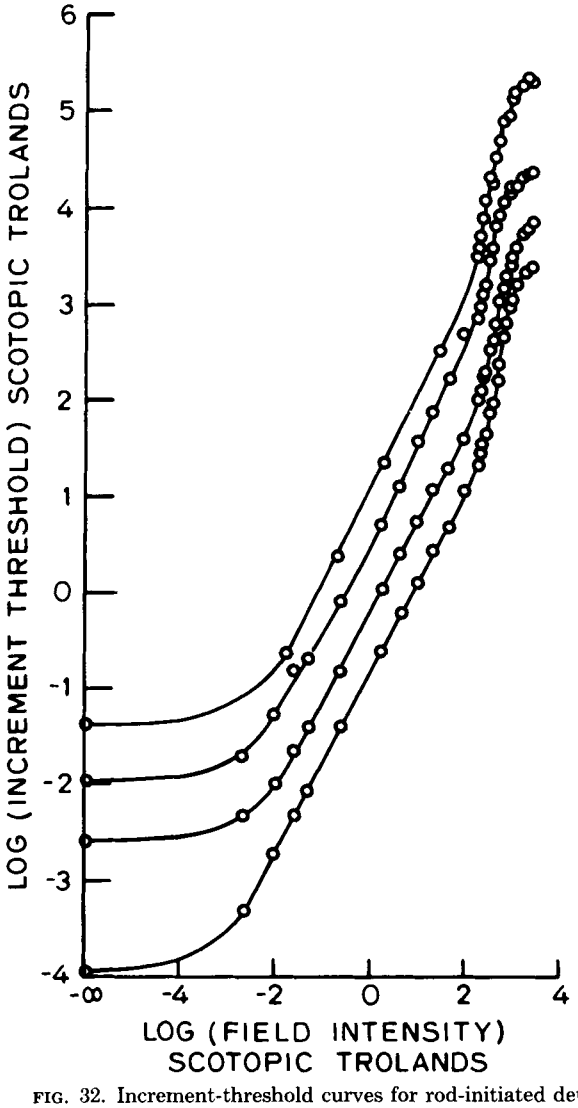

FIG. 32. Increment-threshold curves for rod-initiated detection. Test stimulus, 9° diam, 0.2 s, centered 9° from fovea. Conditioning field, 20° diam, exposed steadily. Means from four observers are shown. *Lowest curve* correctly placed with respect to both axes. Other curves displaced upwards by 0.5, 1.0, and 1.5 log units, respectively. [From Aguilar and Stiles (1a).]

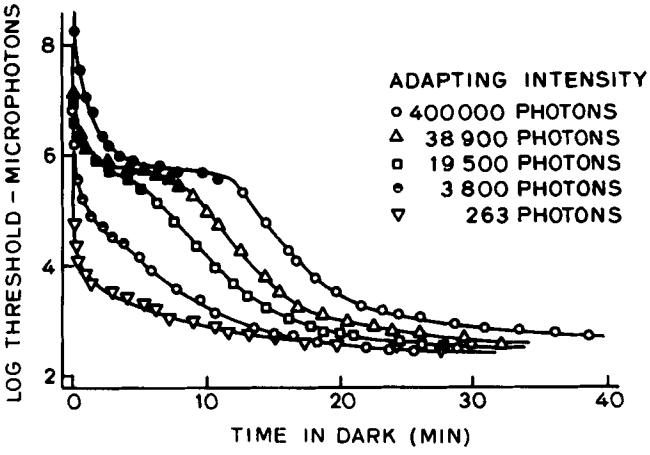

FIG. 33. Dark-adaptation curves follow adaptation to conditioning fields of various intensities. *Filled symbols* indicate thresholds where the test flash had a colored appearance [From Hecht et al. (65).]

To see why this is so, consider the following: If one rejects the hypothesis that sensitivity depends on the quantum catch of a single class of photoreceptors, one can still maintain that a single variable, computed from the quantum catch of all photoreceptors at a proximal level, governs sensitivity (69, 70). Conversely, if one accepts the hypothesis that sensitivity depends on the quantum catch of a single class of photoreceptors, the possibility remains that the state of adaptation in a small region of the retina cannot be characterized by a single real number. An important model that incorporates this feature is a version of the multiple channels model proposed by Campbell and Robson [(34); see also ref. 54]. The model assumes that any region of the retina contains many spatially overlapping, nonidentical channels, each receiving input from only one type of photoreceptor. If the channels adapt independently, sensitivity to a test flash would depend on both *1*) the receptor quantum catch and *2*) the identity of the detecting channel. Thus, no single-variable theory would suffice to describe the state of adaptation, despite the dependence of adaptation on only a single class of photoreceptors.

Because of the logical independence of these issues, we treat them in different sections. We turn now to the identification of neural substrates.

DETECTION BY RODS. Three fundamental psychophysical techniques exist for distinguishing rod-initiated detection from cone-initiated detection.

One method is to measure the observer's sensitivity to different wavelengths of a test light. Since the rods all contain a single pigment, rhodopsin, the wavelength sensitivity to test lights must be equal to the spectral sensitivity of this pigment (corrected for the stable preretinal filters of the eye).

A second method is to measure the change in the observer's threshold as the angle of incidence of the test light is varied. Stiles and Crawford (122) were the first to demonstrate that cone photoreceptors are selectively sensitive to the angle of incidence of light, being maximally sensitive to light passing through the center of the pupil and progressively less sensitive as the bundle of rays pass closer to the pupil's edge. The directional sensitivity of the cones is called the Stiles-Crawford effect of the first kind (39). Rods, however, show no such directional sensitivity (i.e., for angles of incidence within the physical limits imposed by the pupil, but see ref. 48). Therefore, the presence or absence of directional sensitivity is diagnostic of detection based upon a cone- versus rod-initiated signal.

A third method is to use very weak test lights in the periphery of the eye where rods are dense and cones are sparse. Figure 33 shows an example of the way in which threshold declines in the periphery following adaptation to various conditioning fields. The two branches of the recovery curve following the intense conditioning fields have markedly different properties with respect to test-spectral sensitivity and test-directional sensitivity. The upper branch shows directional sensitivity and a spectral sensitivity quite different from rhodopsin. The lower branch shows no directional sensitivity, and test-spectral sensitivity is that of rhodopsin (appropriately corrected). For these and other reasons (see ref. 2) the upper branch is thought to be detection by cones and the lower branch detection by rods. If rod signals do not affect the sensitivity of cone-initiated pathways, the plateau of the recovery curve represents the absolute threshold of cone photoreceptors to the test light. By using a test light below the intensity level of the plateau one renders the test light invisible to cones and thereby forces detection by rods.

Use of these techniques has led to the discovery of other properties that generally, but not reliably, distinguish rod-initiated from cone-initiated detection. For example, scotopic spatial and temporal pooling is generally greater than photopic pooling (8). While these observations are of considerable interest, they are not diagnostic of rod-initiated detection, since under some adapting conditions the spatial and temporal pooling of the scotopic and photopic detection are similar. We refer the reader to Alpern (2) and Barlow (10) for a more complete review of the differences between rod-mediated and cone-mediated detection.

ADAPTATION BY RODS. A strong simplifying hypothesis for scotopic adaptation is the proposal that the quantum catch of only rods governs the state of adaptation. We call this the rod-independence hypothesis. An early and influential study by Flamant and Stiles (48) tested this hypothesis in two ways. Using a large (10°) conditioning field and a square 1° test flash, Flamant and Stiles (48) measured *1*) the action spectrum of the conditioning field for a 1 log-unit elevation of test threshold and *2*) the effect on threshold of varying the angle of incidence of the conditioning field.

The predictions of the rod-independence hypothesis are as follows. First, the action spectrum of the conditioning field must be that of rhodopsin (corrected

for preretinal filters). (A methodologically superior test—used by Flamant and Stiles (48)—is to compare the action spectrum of the conditioning field with the absolute spectral sensitivity of the observer's eye for rod-initiated detection. This comparison is preferred because it obviates problems that may arise due to individual differences in preretinal absorptions.) Second, the effect of the conditioning field must be independent of the angle of incidence of the bundle of rays of the field. This follows from the rod-independence hypothesis because if *1)* adaptation is due to a signal mediated by rods, and *2)* rods show no directional sensitivity, then changes in the angle of incidence of the conditioning field should not affect threshold. Flamant and Stiles's (48) data were consistent with both of these predictions, and they concluded that under their testing conditionings scotopic adaptation was controlled by the quantum catch of the rod photoreceptors, i.e., scotopic sensitivity was independent of cones, and the rod-independence hypothesis could not be rejected.

One limitation of the validity of the rod-independence hypothesis has been reported by Makous and Boothe (92) and Makous and Peeples (93). Using stimulus conditions nearly identical to those of Flamant and Stiles (48)—most importantly a large, 10° background—Makous and Boothe (92) observed differences in the action spectrum of the conditioning field measured at high- and low-criterion levels. When the conditioning field elevated the test threshold about a log unit, or less, they confirmed Flamant and Stiles's findings of an action spectrum that coincided with test sensitivity at absolute threshold. For higher conditioning field intensities, however, which raised threshold about one-and-a-half to two log units, the action spectrum for the conditioning field deviated from test sensitivity at absolute threshold. Moreover, at high field intensities the conditioning field showed a directional sensitivity effect, a finding that was again inconsistent with the rod-independent hypothesis.

Other studies of adaptation have also shown that the independence of the scotopic pathway from the influence of cones obtains only within a limited range of experimental conditions. For example, Lennie and MacLeod (86), Blick and MacLeod (24), Frumkes and Temme (49), and Latch and Lennie (83) measured sensitivity to scotopic test lights using conditioning fields of various diameters. They found that when field diameters are large, as in the Flamant and Stiles (48) experiments, the rod-independence hypothesis cannot be rejected. For small-diameter conditioning fields, however, the action spectrum of the conditioning field differs significantly from the action spectrum for detection under rod-isolation conditions at absolute threshold. This is true even for relatively low intensity conditioning fields. This difference in spectral shape is attributed to the effect of cone photoreceptor signals on the rod-initiated detection pathway.

Lennie and MacLeod (86, 90) suggest a hypothesis to account for these findings. They point out that, in general, the sensitivity measured by a test stimulus falling on one part of the retina will depend on the conditioning-field intensity at retinal locations other than the test region (e.g., ref. 113). Therefore, the pathway that determines psychophysical sensitivity must include neural elements whose sensitivity depends, in part, on spatial pooling across the retina. The distances over which this pooling occurs are quite large, which suggests that a pooling process more extensive than simple receptor coupling is involved.

Lennie and MacLeod (86, 90) thus argue that postreceptor neural adaptation plays a role in determining psychophysical sensitivity. For postreceptor elements, large uniform fields are quite ineffective stimuli. Thus, any part played by postreceptor adaptation should be small when a large, uniform conditioning field is used. As the conditioning-field diameter is reduced toward the size of the test spot, the conditioning stimulus becomes potent for the same class of proximal neurons that are most sensitive to the test spot because of their receptive-field size. Therefore, small conditioning fields should cause relatively more adaptation than large conditioning fields at the proximal neurons that mediate detection of the test spot. Since proximal neurons receive signals from more than a single class of receptors (e.g., refs. 46, 56), sensitivity depends on more than a single class of receptors. Thus, the rod-independence hypothesis is rejected for small conditioning fields, but not necessarily for large fields.

SUMMARY. The isolation of rod-initiated detection pathways has met with considerable success. Identification of such pathways may be made from measurements of test-spectral sensitivity, test-directional sensitivity, and test-intensity levels below the cone plateau of the recovery curve.

The identification of the photoreceptor classes that control the sensitivity of rod-initiated detection pathways has proved subtle. Under certain viewing conditions (e.g., those of Flamant and Stiles, ref. 48) the hypothesis that adaptation is governed by the rod quantum catch alone has not been rejected. Variation, however, in the spatial properties of the backgrounds, or measurements with conditioning fields above one scotopic troland have led to the rejection of the hypothesis. Thus we draw the inference that cone photoreceptors may influence the sensitivity of rod-initiated detection pathways.

Single-Variable Theories

EQUIVALENT-BACKGROUND PRINCIPLE. Stiles and Crawford (121) made the following important argument in 1933. Suppose that the state of adaptation in a small patch of the retina, under steady-state adaptation to conditioning field, μ_1, may be described by a mechanism whose sensitivity is completely characterized by a single, real number. This number might represent the concentration of a chemical substance,

or the number of sites available at a membrane. Threshold at this patch of the retina may be measured with a test probe, λ. The threshold value is a single, real number. If we assume that the threshold of λ is monotonically related to the real number which characterizes the state of adaptation caused by μ_1, then we may take the threshold of λ as a measure of the state of adaptation. Further, should it be the case that λ has the same threshold on conditioning field, μ_2, as on conditioning field, μ_1, we would conclude that the state of adaptation due to μ_2 was the same as that due to μ_1, i.e., μ_2 and μ_1 are equivalent backgrounds, as measured by λ.

Now, if the adapted state of this patch of retina is equivalent under μ_1 and μ_2 adaptation, then threshold to a second test, λ', must also be identical when measured on μ_1 and μ_2. If these thresholds were not identical, we would have to conclude that μ_1 and μ_2 had different adapting effects in contradiction to the measurements with λ. We call the predicted equivalence of these backgrounds (for all test flashes) the *equivalent-background* principle.

Should empirical conditions exist where μ_1 and μ_2 are equivalent with respect to a first test, λ, but not with respect to a second test, λ', we would have to reevaluate some part of our hypothesis. One difficulty may be that λ and λ' are measuring the state of adaptation of different physiological pathways, although the state of each pathway can be well described by a single variable. Alternatively, λ and λ' may be detected by the same anatomical pathway, but the single-variable hypothesis is false; that is, the state of adaptation may require more than one variable in order to be completely described.

TESTS OF EQUIVALENT-BACKGROUND PRINCIPLE. With these arguments in mind, Stiles and Crawford (121) carried out an experimental test of single-variable theories of light adaptation. They used three kinds of conditioning fields: *1*) a glare source (point of light) 6° above the observer's fixation, *2*) an annulus with inner diameter of 1° and outer diameter extending over the whole visual field, and *3*) a uniform field, also extending over the entire visual field. For each of these conditioning fields, Stiles and Crawford (121) determined the intensity required to elevate six test lights of various spatial, temporal, and spectral parameters to an equal level. This was done with both foveal and parafoveal presentation of the stimuli, and for two intensity levels of the test light.

There were measurable deviations from the equivalent-background principle of the single-variable hypothesis. The data are somewhat problematic, however, because Stiles and Crawford (121) made no attempt to isolate different visual mechanisms in their measurements. Thus failures of the equivalent-background principle of single-variable theories could have been due to either differences in the pathways being probed, or a failure of the single-variable hypothesis.

Crawford (38) later returned to the equivalent-background principle. He tested the single-variable hypothesis using experimental conditions different from those reported by Stiles and Crawford (121). Specifically, Crawford (38) extended the equivalent-background measurement to include the dynamic case.

The logic of Crawford's (38) extension, also described in Stiles and Crawford's (121) 1933 paper, is as follows. For a test, λ, threshold is measured at various intensities of a conditioning field, μ_1. Threshold recovery is then measured following adaptation to a bright conditioning field, μ_2, using this same test, λ. For each intensity of μ_1, a time during recovery from μ_2 is found such that λ threshold at that intensity of μ_1 and at that time of recovery from μ_2 were equal. This generated a set of equivalences. Each intensity of μ_1 is equivalent to some time during recovery from μ_2. On a single-variable theory, when adaptation is measured with a different test, λ', the various intensities of μ_1 must remain equivalent to the same set of times during recovery.

Crawford (121) performed such measurements with six test stimuli. His results did not reject the equivalent-background prediction. Thus, the single-variable hypothesis was consistent for equivalences between light and dark adaptation but inconsistent for the equivalences between different spatial distributions of steady-state conditioning fields (121).

Again, however, Crawford's (121) data are problematic. One problem is that fixation was not carefully controlled. (His data were taken for applied reasons during World War II.) A second problem is that the separation between rod- and cone-initiated detection is not at all clear from his measurements. Hence it is difficult to know which physiological mechanisms are being investigated.

Because of the general importance of single-variable theories, the problem was again studied by Blakemore and Rushton (22). The subject in their experiments was a rod monochromat, so the problem of isolating a rod-initiated detection pathway was solved by nature. The equivalent-background principle, derived from single-variable theories, was tested following Crawford's method of comparing increment thresholds and dark-adaptation curves. In Figure 34 we have replotted Blakemore and Rushton's (22) measurements. On the right are two increment-threshold curves, for test lights of 5′ and 360′ upon a 3,600′-diameter conditioning field. On the left-hand side are two recovery curves for these same test lights. The conditioning field was again the large, 3,600′ field, at an intensity that Blakemore and Rushton (22) estimated to bleach "about half" of the rhodopsin in the rods.

These recovery curves illustrate an important observation emphasized by earlier writers including Craik and Vernon (37), Lythgoe (88), and Arden and Weale (6). Notice that the vertical separation between the recovery curves for the small and large test flashes grows as recovery proceeds. The increasing separation

indicates that the eye is better able to integrate rod signals across the retinal surface as recovery proceeds. This increase in spatial integration is accompanied by a decrease in spatial acuity (22, 30). These observations are difficult to reconcile with any theory that supposes the photopigment concentration as the direct mechanism that limits sensitivity. Rather, some sort of neural changes appear to be involved (37, 88).

As Rushton, Barlow, and their collaborators (7, 9, 10, 110, 111, 113) point out, however, it remains possible that the nonphotochemical mechanisms are controlled by a signal whose value is determined exclusively by the photopigment concentration. Thus, while postreceptor mechanisms may be involved in the process of adaptation, they may all be driven by a single common variable. In such a circumstance, the equivalent-background experiment still measures equivalences in rhodopsin concentration (a one-dimensional variable) and thus the equivalent-background principle should still obtain.

The dotted lines in Figure 34 illustrate the test of the equivalent-background principle for these increment-threshold and recovery curves. For the 5′-flash threshold, 15 min into recovery was equal to threshold with a steady conditioning field at approximately 4 scotopic trolands. Thus the equivalent-background principle asserts that 15 min into recovery and a 4 scotopic-trolands background should generate an equal threshold for any other test. Blakemore and Rushton (22) tested this prediction for the 360′-test flash shown in the figure. The agreement here, as well as for other time and intensities, is very good.

On the basis of these (and other) observations, Rushton (111) concluded that the single-variable hypothesis could not be rejected. Using the letter G (gain) as the variable defining adaptation, he wrote:

> ... what Crawford did was to change G by bleaching [dark adaptation] and by backgrounds [increment thresholds] and to compare the organization of the G-box in the two cases using different types of test ΔI. He found that when the G settings matched by one test they matched by all. Intensity and spatial integration were not independent variables that needed separate matching adjustments. All the factors that G controls are the function of a single variable (one knob).

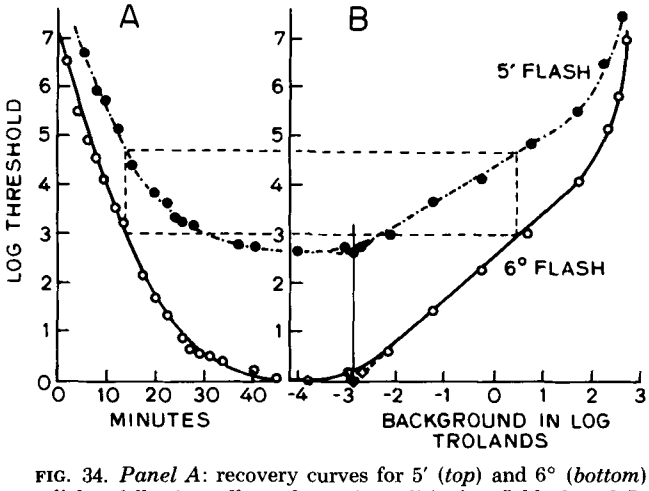

FIG. 34. *Panel A*: recovery curves for 5′ (*top*) and 6° (*bottom*) test lights following offset of 3,600′ conditioning field. *Panel B*: increment-threshold curves for the same test lights superimposed on a 3,600′ steady-background field. [From Blakemore and Rushton (22).]

A SPECIFIC SINGLE-VARIABLE THEORY: THE NOISE HYPOTHESIS. An important, single-variable hypothesis was studied by Barlow (7, 9). Following the suggestion of Rose (107) and de Vries (44), Barlow tested the hypothesis that threshold elevation on intense conditioning fields is due to noise in the receptor signals. According to this view, the single-variable controlling psychophysical sensitivity is the level of receptor noise. In the light-adaptation experiment the inevitable quantal fluctuations of the bleaching process generate the noise. In the dark the time-varying free opsin produced by photopigment bleaching generates the receptor noise.

In addition to tests that apply to all single-variable theories of adaptation, the receptor-noise hypothesis, together with the assumption of optimal signal-to-noise discrimination, leads to the prediction that threshold will rise as the square root of the conditioning-field intensity. This prediction is generally contradicted when long-duration, large-area test spots are used (e.g., see refs. 1a, 8). Increment-threshold curves for small, brief tests follow a different increment-threshold curve from long, large tests. The slope of this curve grows more slowly. Thus increment thresholds for small, brief stimuli can be tolerably well fit over a somewhat greater range by a curve obeying the square-root relationship than can increment-threshold curves for large, long-duration test stimuli. This has led some authors to suggest that the theory may be correct for a restricted range of test stimuli (e.g., ref. 108).

The receptor-noise hypothesis, however, as an explanatory principle is only of restricted value. For, as Stiles wrote (120):

> The reason why a uniform field raises the increment threshold is now often discussed in relation to the inevitable fluctuations in the number of light quanta from the field that are absorbed by the individual receptors, or strictly by completely summating groups of receptors. The test stimulus intensity must somewhat exceed these fluctuations, if it is to be detected. This is certainly true. But since we do not in fact see the fluctuations as such we must also suppose that somewhere in the neural chain a barrier is raised high enough to prevent the fluctuations passing into consciousness. The raising of this barrier corresponds to an increase [*sic*: decrease is meant] in the effective sensitivity of the receptors of summating receptor groups. It would seem that this is the true adaptational change. The fluctuations theory merely places a lower limit on the height of the barrier: its nature remains to be explained.

RELATED STUDIES. Further work on the problem of equivalent backgrounds was reported by Barlow and Sparrock (14) and Westheimer (139).

Barlow and Sparrock (14), working from Barlow's receptor-noise hypothesis, asked why the noise signal during dark adaptation did not appear as bright as the equivalent-background noise signal in the light-adaptation experiment. They tested the hypothesis that the brightness difference between the afterimage (7°) and real-light (28°) background that raised threshold by equal amounts was due to the stabilization of the afterimage. They measured the apparent brightness of the afterimage by matching its appearance with a stabilized annulus. The intensity of the stabilized annulus that matched the afterimage in brightness was equal to the intensity of the real-light background that raised the test threshold by the same amount as the afterimage. This provided further support for a single-variable theory but not particularly for the receptor-noise hypothesis.

Westheimer (139) examined the following important point. Suppose we perform an equivalent-background experiment with conditioning field, μ_1, and bleaching field, μ_2, that have the same spatial distribution on the retinal surface. This permits identification of intensities of μ_1 as being equivalent to times during recovery from μ_2.

Now suppose we vary the spatial distribution of μ_1 and μ_2 to a new, but common, spatial distribution using the same test light. The question posed by Westheimer (139) is: Will the set of equivalences of intensities and times—as measured with the same test λ—remain unchanged as the spatial distribution of the background is varied? The answer is no. Rather, varying the spatial distribution of the conditioning fields generates a new set of equivalences.

These observations are not inconsistent with the single-variable hypothesis. But what must be true on the single-variable hypothesis—and is so far untested—is that these new equivalences are preserved as the test flash parameters are varied. The general issue of equivalence and some potential problems in the experimental methodology of Westheimer's work is reviewed by Barlow and Sakitt [(13); see also ref. 123].

A CONSTRAINT ON SINGLE-VARIABLE THEORIES: THRESHOLD RECOVERY. Historically, the most important hypothesis to account for threshold recovery following adaptation to a conditioning field has been the effort to equate the state of adaptation with the concentration of visual pigment, rhodopsin, in the observer's rod outer segments. For a general chemical light-driven reaction, we may write that the change in concentration, s, over time, t, depends on the light intensity, I, via

$$\frac{ds}{dt} = -I\phi_1(s, x_1, x_2, \ldots, x_n) + \phi_2(s, x_1, x_2, \ldots, x_n)$$

where the x_i (i.e., $x_1, x_2, \ldots, x_n$) are variables that affect the reaction. We now know from Rushton's (109) and Weale's (134–136) studies of the bleaching of human rhodopsin in situ that pigment kinetics obey first-order equations quite well. Thus

$$\frac{ds}{dt} = -I\phi_1(s) + \phi_2(s)$$

Consider the hypothesis that sensitivity depends monotonically only on the instantaneous pigment concentration in the observer's eye, s. Following equilibrium bleaching by a conditioning field, s will vary as

$$\int_{s_0}^{s} \frac{ds}{\phi_2(s)} = t - t_0$$

Letting F be the (monotonic) indefinite integral of ϕ_2, it follows that

$$s = F\sum^{1}[t - (t_0 - F(s_0))] \quad (1)$$

Thus, this hypothesis and first-order kinetics of bleaching lead to the conclusion that all recovery curves following the offset of a conditioning field must follow the same time course, up to displacement with respect to the time axis by an amount $t_0 - F(s_0)$ (Eq. 1).

This argument was noted by Winsor and Clark (143) as well as Lythgoe (88), Crawford (38), and Stiles (120). Figure 35 replots Winsor and Clark's (143) measurements of scotopic recovery to a 1.3°, 20-ms test following adaptation to three levels of a large conditioning field. Clearly, no sliding of these curves along the time axis will bring their rod branches into coincidence.

Furthermore, recall that Figure 33 replots the results of Hecht et al. (65). Again, it can be seen that no horizontal displacement of these curves will bring the rod branches into coincidence.

Finally, we note that Pugh (100, 101), in a recent study of rod dark adaptation and rhodopsin regeneration in situ has shown that rod dark-adaptation curves, following adaptation to different intensity conditioning lights, follow different time courses. In Figures 36*A* and 36*B* we replot Pugh's data showing the relationship between the time constant of the best-

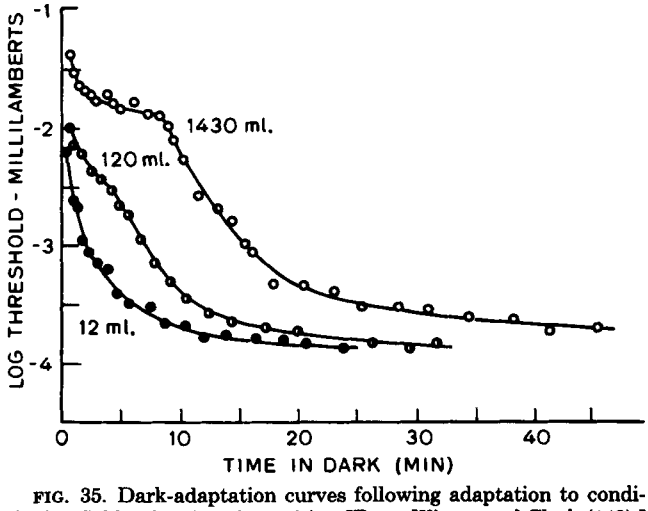

FIG. 35. Dark-adaptation curves following adaptation to conditioning fields of various intensities. [From Winsor and Clark (143).]

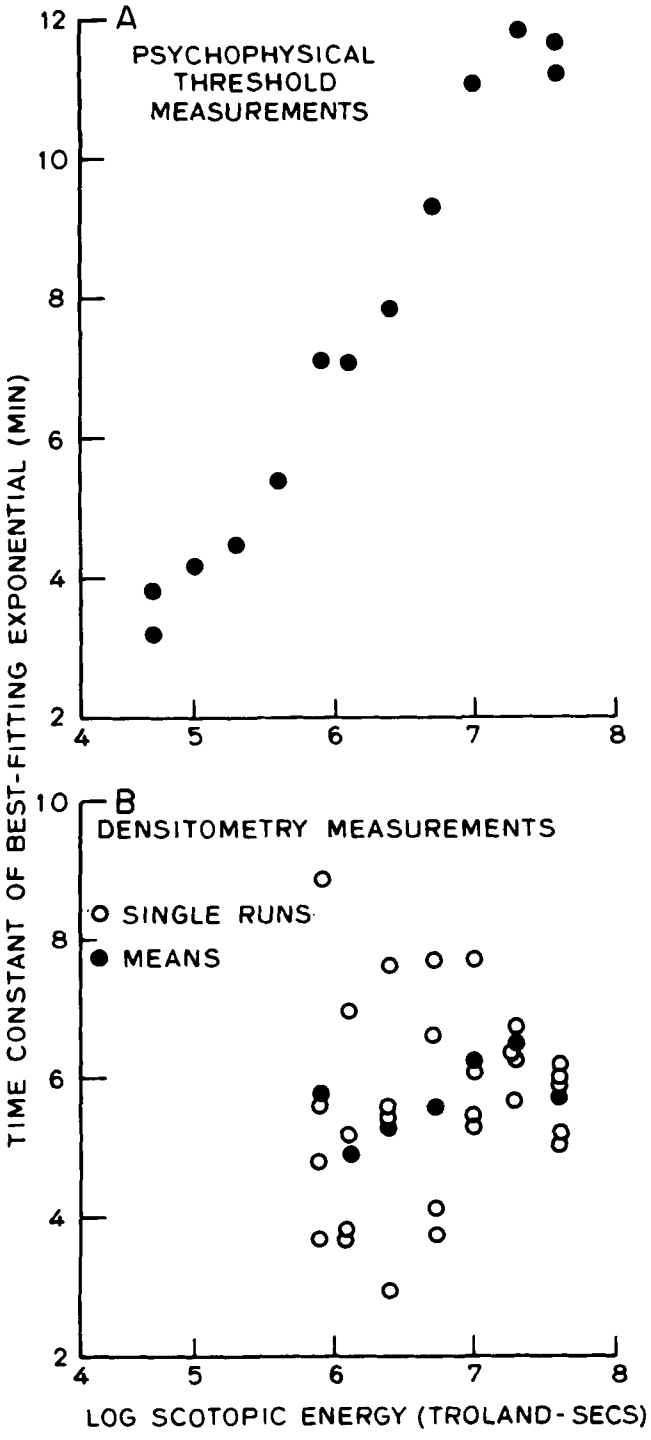

FIG. 36. Time constant of best-fitting exponential curve to *A*: psychophysical threshold recovery during dark adaptation, and *B*: rhodopsin regeneration in the living human eye. Both are plotted as a function of energy in the conditioning field. *Open symbols* in *B* are individual runs, *filled symbols* are the arithmetic means of individual runs. [Data from Pugh (100, 101).]

fitting exponential curve to threshold recovery (Fig. 36*A*) and rhodopsin regeneration (Fig. 36*B*) as a function of conditioning-field energy. Notice that the time course of rhodopsin regeneration is approximately constant, independent of bleaching energy. Threshold recovery, however, proceeds more slowly following more intense bleaches.

Winsor and Clark (143) concluded from the failure of their prediction of a single shaped curve that either *1*) first-order kinetics equations were false, or *2*) sensitivity depended on variables other than the fraction of bleached pigment. Since no objective measurements of rhodopsin kinetics were then available, and Wald's (128, 129) work had led them to believe that rhodopsin kinetics would not be first order, Winsor and Clark (143) argued that *1*) and not *2*) should be rejected.

Twenty years after Winsor and Clark's (143) experiments, Rushton (109) and Weale (134) accomplished the difficult feat of measuring human rhodopsin in situ. From their measurements it became clear that rhodopsin density did follow first-order pigment kinetics following equilibrium adaptation. Thus, since *1*) is true within measurement error, it follows that *2*), the dependence of threshold solely upon rhodopsin concentration, must be rejected.

Notice that the objectionable aspect of the hypothesis is the identification of bleached rhodopsin as the controlling factor. The results do not allow us to reject all single-variable theories. However, Winsor and Clark's (143) arguments demand that if sensitivity depends on a single substance, that substance cannot obey first-order pigment kinetics. It is possible, however, that there exists a controlling substance, *S*, and that its concentration (*s*) depends on at least two variables. In such a case, the time course of recovery for this substance would have the functional form

$$s = H[\psi_1(t),\psi_2(t)]$$

where ψ_1 and ψ_2 are the two variables on which the concentration, *s*, depends. This single-variable theory—where the controlling variable's recovery depends on two variables—is the simplest form to which we must advance.

SUMMARY. The time course of dark adaptation, at one time thought to be identified with rhodopsin concentration, depends on more than a single variable. If a single substance is found whose concentration is perfectly correlated with the sensitivity of rod-initiated pathways, that material's concentration must depend on the effect of at least two variables (120, 143). We use the term *substance* broadly to include variables such as membrane sites.

Concluding Comment

The single-variable hypothesis tested by the equivalent background postulated the existence of some unnamed substance, or site, whose value completely characterizes the state of adaptation. The classic view of adaptation as a process controlled by the concentration of photopigment—whether Hecht's (64) photochemical depletion, Wald's (130) compartments, Rushton's (111) gain-box, or Barlow's (9) dark noise—is currently being questioned. A search for new substrates and mechanisms must follow. The advantage

of testing the equivalent-background principle is that its failure would signify the failure of all single-variable theories. Hence, no single substance or site could suffice to describe the state of adaptation. If the equivalent-background principle were accepted over a wide range of experimental conditions, however, the hope of identifying a single controlling substance would increase.

NOTE ADDED IN PROOF

Since this chapter was completed, new developments have suggested answers to both of the psychophysical puzzles mentioned in *Concluding Comment*, p. 300. Publishing logistics forbid a thorough discussion of these developments—the reader is referred instead to the references at the end of this note. However, the highlights can be briefly outlined.

First, it now appears that the nearly unanimous failure to detect foveal interference fringes above 60 cycles/deg has been a matter of psychophysical methodology rather than retinal physiology. Using a refined interferometer, D. R. Williams (personal communication) has recently shown that spatial frequencies at least as high as 130 cycles/deg can definitely be detected in the foveola. The appearance of gratings at these very high frequencies seems entirely consistent with aliasing by the cones, whose inner segments in the foveola approximate a spatially regular hexagonal lattice. [Fig. 18 in this chapter shows an array of foveal outer segments. Miller and Bernard (3A) have recently presented a convincing optical argument that the image-sampling properties of photoreceptors should be analyzed on the assumption that the effective aperture of the receptors is located midway along their inner segments. Miller (2A) and Borwein et al. (1A) have shown that in the rod-free central fovea the spatial arrangement of the inner segments is much more regular than that of the outer segments. Thus our Fig. 18 suggests more spatial disorder than is apparently present in the effective foveal image sampling array.]

Second, it now appears that aliasing outside the fovea in normal vision is forestalled by spatial disorder in the photoreceptor array. Nagel (4A) pointed out that the consequences of image sampling by a spatially irregular array, such as the extrafoveal cones, are determined by the Fourier power spectrum of the array. Mathematically, the key fact is that the power spectrum of any image after sampling by a random array is its original power spectrum convolved with that of the sampling array. Subsequently Yellott (5A) computed the power spectra of sections of rhesus cones ranging across the entire retina. These spectra reveal that throughout the extrafoveal retina the cones provide a form of optimal random sampling—spatial frequencies above the aliasing cutoff implied by local cone density are not aliased back to discrete lower frequencies as would happen with spatially regular sampling arrays, but instead they are scattered into broadband spatial noise at all orientations. On the other hand, spatial frequencies below the nominal aliasing cutoff (which could have been recovered with no distortion by the same number of cones arranged in a regular lattice according to the classical sampling theorem) are transmitted with minimal noise. These twin goals are accomplished by a spatial arrangement based on constrained random placement [Yellott (6A)].

Overall, it now appears that in normal photopic vision aliasing is prevented in the fovea by the optics of the eye, which bandlimit the retinal image to 60 cycles/deg, and outside the fovea by an optimal spatial disorder in the cone mosaic.

1A. BORWEIN, B., D. BORWEIN, J. MEDEIROS, AND J. W. MCGOWAN. The ultrastructure of monkey foveal photoreceptors, with special reference to the structure, shape, size, and spacing of the foveal cones. *Am. J. Anat.* 159: 125–146, 1980.

2A. MILLER, W. H. Intraocular filters. In: *Handbook of Sensory Physiology. Invertebrate Photoreceptors*, edited by H. Autrum. Berlin: Springer-Verlag, 1979, vol. 7, pt. 6A.

3A. MILLER, W. H., AND G. D. BERNARD. Averaging over the foveal receptor aperture curtails aliasing. *Vision Res.* In press, 1983.

4A. NAGEL, D. C. Spatial sampling in the retina. *Invest. Ophthalmol.* 20, Suppl. 3: 123, 1981.

5A. YELLOTT, J. I., JR. Spectral consequences of photoreceptor sampling in the rhesus retina. *Science* 221: 382–385, 1983.

6A. YELLOTT, J. I., JR. Nonhomogeneous Poisson disks model the photoreceptor mosaic. *Invest. Ophthalmol.* 24, Suppl. 3: 145, 1983.

NOTE: The complete list of references for this chapter and its *Appendix* starts on p. 313.

Appendix: Fourier transforms and shift-invariant linear operators

CONTENTS

THIS APPENDIX OUTLINES the basic mathematical ideas underlying the kind of Fourier analytic treatment of retinal imagery given earlier in VISUAL ACUITY, p. 260. Broadly speaking the idea is that inputs to the eye can be thought of as functions [i.e., $f(x,y)$ represents the light intensity in the object scene at spatial point (x,y)]; associated with each input f is an equivalent function, t_f (the *Fourier transform* of f), which describes f as a weighted sum of sinusoidal components. The effect of passing any input through the optics of the eye is then represented by an operator, O, that transforms each input function, $f(x,y)$, into an output function, $O[f(x,y)]$, corresponding to light intensity in the retinal image of f. If O satisfies two conditions (*linearity* and *shift-invariance*, both assumed to be approximately valid for the optics of the eye), its effect can be completely specified either by an associated function $o(x,y)$, called the *impulse response*, or by the Fourier transform of o, t_o, which is called the *transfer function* of O. If either of these functions is known, the retinal image corresponding to any input (or equivalently, its Fourier transform) can be immediately written down. Consequently if one knows either the impulse response, $o(x,y)$, or the transfer function, t_o, of an eye, one can calculate the retinal image corresponding to any visual stimulus and thereby determine what information is actually available at the level of the photoreceptors.

The purpose of this section is to flesh out these ideas by sketching the basic mathematical steps explicitly. Fourier analytic notions have been very widely employed in physiological optics for some time, but we are not aware of any text devoted specifically to this application. Bracewell (27) is an excellent practical introduction to the transform generally, Goodman (51) and Hecht and Zajac (62) cover Fourier optics, and Lighthill (87) explains the mathematical justification for the convenient operational calculus methods which have made the Fourier transform such a popular analytic tool.

ONE-DIMENSIONAL CASE

Although optical problems generally deal with functions of two variables, e.g., x and y, corresponding to the horizontal and vertical axes of an image plane, it is more convenient mathematically to begin with functions of a single variable. Suppose $f(x)$ is a real-valued function of a real variable x, for example, $f(x)$ might represent a light-intensity profile along the x axis of the retinal plane. The Fourier transform of f is another function, $t_f(s)$, given by the integral

$$t_f(s) = \int_{-\infty}^{\infty} e^{-i2\pi sx} f(x)\, dx \qquad \text{(A1)}$$

where e is the constant 2.71828 . . . and i is the imaginary unit $\sqrt{-1}$. In general $t_f(s)$ is a complex function of the real variable s: That fact can be made explicit by using the Euler relationship $e^{i\theta} = \cos\theta + i \sin\theta$ to rewrite Equation A1 in the form

$$\begin{aligned} t_f(s) &= \int_{-\infty}^{\infty} f(x) \cos 2\pi sx\, dx \\ &\quad - i \int_{-\infty}^{\infty} f(x) \sin 2\pi sx\, dx \qquad \text{(A2)} \\ &= \text{Re } t_f(s) + i \text{ Im } t_f(s) \end{aligned}$$

Here Re $t_f(s)$ and Im $t_f(s)$ are the real and imaginary parts of t_f; both are real functions of s. Recalling that any complex number $z = a + ib$ can be written in the form $z = |z|\, e^{i \text{ pha}(z)}$ [where $|z| = \sqrt{a^2 + b^2}$ and pha(z) is the angle through which a vector of length $|z|$ must be rotated to bring its tip into coincidence with the point (a,b)] Equation A2 can be rewritten in the form

$$t_f(s) = |t_f(s)|\, e^{i \text{ pha } t_f(s)} \qquad \text{(A3)}$$

Here $|t_f(s)|$ is a nonnegative real function of s, called the *amplitude spectrum* of f, and pha $t_f(s)$ is another real function of s called the *phase spectrum* of f. The names are motivated by the roles these functions play when $f(x)$ is represented as a weighted sum of sinusoids, as we show next.

The fundamental theorem of Fourier analysis (*Fourier's inversion theorem*, e.g., ref. 87, p. 16) shows that a function f and its transform t_f contain exactly the same information, in the sense that if t_f is known, f can be recovered from the relationship

$$f(x) = \int_{-\infty}^{\infty} e^{i2\pi sx}\, t_f(s)\, ds \qquad \text{(A4)}$$

If $t_f(s)$ is expressed in exponential form (i.e., Eq. A3), the inversion formula (Eq. A4) can be rewritten (after a little calculation) as

$$f(x) = \int_{-\infty}^{\infty} |t_f(s)| \cos[2\pi sx + \text{pha } t_f(s)]\, ds \qquad \text{(A5)}$$

This version shows explicitly that f can be expressed as a weighted sum of sinusoidal components: $|t_f(s)| \cos[2\pi sx + \text{pha } t_f(s)]$ is a cosine of amplitude $|t_f(s)|$ and a frequency of s cycles per unit of x. The phase term pha $t_f(s)$ means that this cosine is shifted to the left along the x axis by an amount equal to $(2\pi s)^{-1}$ pha $t_f(s)$. Consequently Equation A5 expresses the fact that the original function $f(x)$ can be thought of as a sum of cosines of various frequencies: For each frequency s, the amplitude spectrum $|t_f(s)|$ tells how much to weight the cosine term at that frequency, and pha $t_f(s)$ how much to shift it to make the entire sum (integral) equal f. [Actually, the amplitude of the cosine at frequency s is $2|t_f(s)|$ because the integral (Eq. A5) contains terms at $+s$ and $-s$; each represents the same cosine component and each has amplitude $|t_f(s)|$.] Notice that we cannot generally expect to recover a function from its amplitude spectrum alone; the phase spectrum also must be known.

Generalized Functions

It is apparent from the defining Equation A1 that many useful functions (such as the constant function $f(x) = 1$, $\cos x$, etc.) do not actually have Fourier transforms because the necessary integration cannot be carried through. Nevertheless such functions can still be treated with Fourier analytic methods. The mathematical justification for this extension is provided by the theory of generalized functions (87), in which transforms of functions such as $\cos x$ are defined in terms of limits of sequences of transforms of functions which converge to $\cos x$. The resulting entities are called transforms in the limit, and for most purposes can be formally manipulated in the same way as ordinary transforms, provided one exercises a little care. For example, the transform (in the limit) of $\cos 2\pi\phi x$ (a cosine of frequency ϕ cycles per unit of x) is a pair of impulses (Dirac delta functions) located at $s = \pm\phi$. These entities can be thought of as very tall, thin rectangular pulses—so thin that their horizontal spread can be treated as zero, even though the total area under the pulse is nonzero. [Bracewell (27) explains this concept in detail.]

Shift-Invariant Linear Operators and Convolution

Suppose $\circ$ is an operator that transforms real functions into real functions; $\circ$ can be thought of as a black box to which we input a function $f(x)$ and observe an output function $\circ[f(x)]$. For example, if $f(x)$ is a light-intensity profile along the x axis of a visual stimulus, $\circ[f(x)]$ might represent the intensity profile along the x axis in the retinal image. In this case the operator $\circ$ represents the degradation of the image imposed by the optics of the eye. The symbol $\circ$ is said to be a linear operator if the output corresponding to the sum of two input functions, f and g [i.e., we input $f(x) + g(x)$], is always the sum of the two outputs produced by f and g separately, i.e., $\circ[f(x) + g(x)] = \circ[f(x)] + \circ[g(x)]$. The symbol $\circ$ is said to be *shift invariant* if the output to $f(x + S)$ [i.e., $f(x)$ shifted bodily S units to the left] is the output to $f(x)$ shifted by the same amount S [i.e., if the output to $f(x)$ is $\circ(x)$, then the output to $f(x + S)$ is $\circ(x + S)$]. Many physically important operators are both shift invariant and linear (*SIL*), or at least satisfy both conditions closely enough for practical purposes.

The beauty of SIL operators depends on two theorems. First, if an operator is SIL it can always be represented as a *convolution*. If $\circ$ is SIL, then there exists a function o such that for every input function f

$$\circ[f(x)] = \int_{-\infty}^{\infty} f(x - r)o(r)\, dr \qquad \text{(A6)}$$

i.e., the output to f is the convolution of f and o (denoted $f * o$). Intuitively this means that the output $\circ[f(x)]$ at any point, x, is a weighted sum of the inputs at $x \pm r$, $0 \leq r < \infty$, each input being weighted by a factor $o(r)$ that depends on the distance from x. Consequently the operator $\circ$ is completely specified by its weighting function o. This function o is called the *impulse response* of $\circ$ because when the input function is a single brief pulse of unit energy delivered at $x = 0$, the output function is exactly $o(x)$. (In optical applications, the one-dimensional impulse response is called the *linespread function*.)

The second important property of SIL operators arises from the fact that the Fourier transform of the convolution of two functions f and g is the product of the transforms of f and g alone, i.e.

$$t_{f*g}(s) = t_f(s)t_g(s)$$

Consequently from a transform standpoint the effect of applying an SIL operator $\circ$ to any input f is simply to multiply the original transform t_f by the transform of the impulse response of $\circ$, i.e.

$$t_{\circ[f]}(s) = t_o(s)t_f(s) \qquad \text{(A7)}$$

This relationship in the transform domain enormously simplifies analysis of SIL operators, and by extension, the analysis of systems in which several such operators are applied sequentially (e.g., suppose an operator O_1 is first applied to an input f, and then a second operator O_2 is applied to the result, so that the final output of the system is $O_2[O_1[f(x)]]$. Then the Fourier transform of the output will be simply the product $t_f(s)(t_o)_1(s)(t_o)_2(s)$. It should be clear that an SIL operator O can be specified either in terms of its impulse response $o(x)$, or by the Fourier transform of $o, t_o(s)$. The latter function is called the *transfer function* of the operator. Using the fact that for any pair of complex numbers z_1, z_2 $|z_1 z_2| = |z_1||z_2|$, it follows immediately from Equation A7 that the amplitude spectrum of the output of an operator, O, applied to an input, f, is the product of the amplitude spectra of the input and the impulse response, i.e.

$$|t_{O[f]}(s)| = |t_f(s)|\,|t_o(s)| \tag{A8}$$

The amplitude spectrum of o, i.e., $|t_o(s)|$, is called the *modulation transfer function* (*MTF*) of O, because it specifies how the amplitude of sinusoidal modulation at each frequency in the input is altered by O. If the input contains a sinusoidal component at a frequency, s, having an amplitude of A (i.e., $|t_f(s)| = A$) the output contains a sinusoidal component at frequency s having amplitude $A\,|t_o(s)|$. Thus if $t_o(s)$ is zero, the operator cannot transmit sinusoidal modulation at frequency s; that information is lost when f is passed through O.

In general an SIL operator alters not only the amplitudes of the sinusoidal components of an input but also their phases. Writing the transform in Equation A7 in exponential form (using Eq. A3) we obtain

$$t_{O[f]}(s) = |t_f(s)|\,|t_o(s)|\,e^{i[\text{pha } t_f(s) + \text{pha } t_o(s)]} \tag{A9}$$

The interpretation of Equation A9 is that the cosinusoidal component in the input at frequency s reemerges in the output shifted to the left on the axis by an amount equal to $(2\pi s)^{-1}$ pha $t_o(s)$. Pha $t_o(s)$ can be regarded as the *phase transfer function*, analogous to MTF. If both $|t_o|$ and pha t_o are known, the operator O is completely specified, since the transfer function, $t_o(s)$, is $|t_o(s)|e^{i\,\text{pha } t_o(s)}$.

EXAMPLES. Figure A1 illustrates the concepts just discussed, with special reference to the phenomenon of spurious resolution produced by defocusing (i.e., the effect demonstrated subjectively by Fig. 11). Across the figure each sequence of four panels represents successively a function, its Fourier transform (in all these cases the transform is entirely real and so can be graphed as a single function), the amplitude spectrum (which in these cases is just the absolute value of the transform graph next to it), and finally the phase spectrum. The impulse function $\delta(x)$, that is, in effect, a rectangle of unit area and zero width (i.e., the Dirac delta function mentioned earlier in this *Appendix*) is represented by Sequence A. A function which provides a very good fit to the one-dimensional impulse response (i.e., the linespread function) of the human eye for a 2-min pupil as determined objectively by Campbell and Gubisch (33) is illustrated in Sequence B. The transform of this function is the transfer function of the eye. The profile of a cosine grating at 100% contrast is illustrated in Sequence C. The effect of convolving that input with the linespread function B is illustrated in Sequence D. Sequence E shows a rectangular window of width b and unit area. The convolution of this function with an input represents the effect of defocusing, i.e., the image of a point is spread over a "blur segment" (analogous to the two-dimensional blur circle) of width b. Finally, the combined effect of defocusing and line spread is illustrated in Sequence *F*. The transform here is the product of the transforms in Sequences B, C, and E. When the input cosine frequency ϕ falls between $1/b$ and $2/b$ the output cosine is reversed in sign so that hills in the input become valleys in the output and vice versa, as shown in the figure.

TWO-DIMENSIONAL CASE

Suppose now that f is a real-valued function of two real variables, e.g., $f(x,y)$ denotes light intensity at the spatial point (x,y). All of the Fourier analytic concepts from the one-dimensional case can readily be generalized to two (or more) dimensions. No new ideas are involved, though naturally the technical apparatus becomes somewhat more complicated. The two-dimensional Fourier transform of f is given by

$$t_f(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i2\pi(ux+vy)} f(x,y)\,dx\,dy \tag{A10}$$

which is a complex function of two real variables, u and v. The *inversion formula* becomes

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i2\pi(ux+vy)} t_f(u,v)\,du\,dv \tag{A11}$$

In the same way that the one-dimensional case (Eq. A4) can be rewritten in a more informative version (Eq. A5), Equation A11 can be rewritten in a form that shows explicitly how $f(x,y)$ is composed of a weighted sum of elementary sinusoidal functions

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |t_f(u,v)| \cos[2\pi(ux + vy) + \text{pha } t_f(u,v)]\,du\,dv \tag{A12}$$

Here $|t_f(u,v)|$ is again the amplitude spectrum of f, pha $t_f(u,v)$ the phase spectrum. The function $\cos[2\pi(ux + vy) + \text{pha } t_f(u,v)]$ describes a vertical cosinusoidal grating in the x,y plane [i.e., a two-dimensional function of the form $f(x,y) = \cos 2\pi\phi x$ as shown

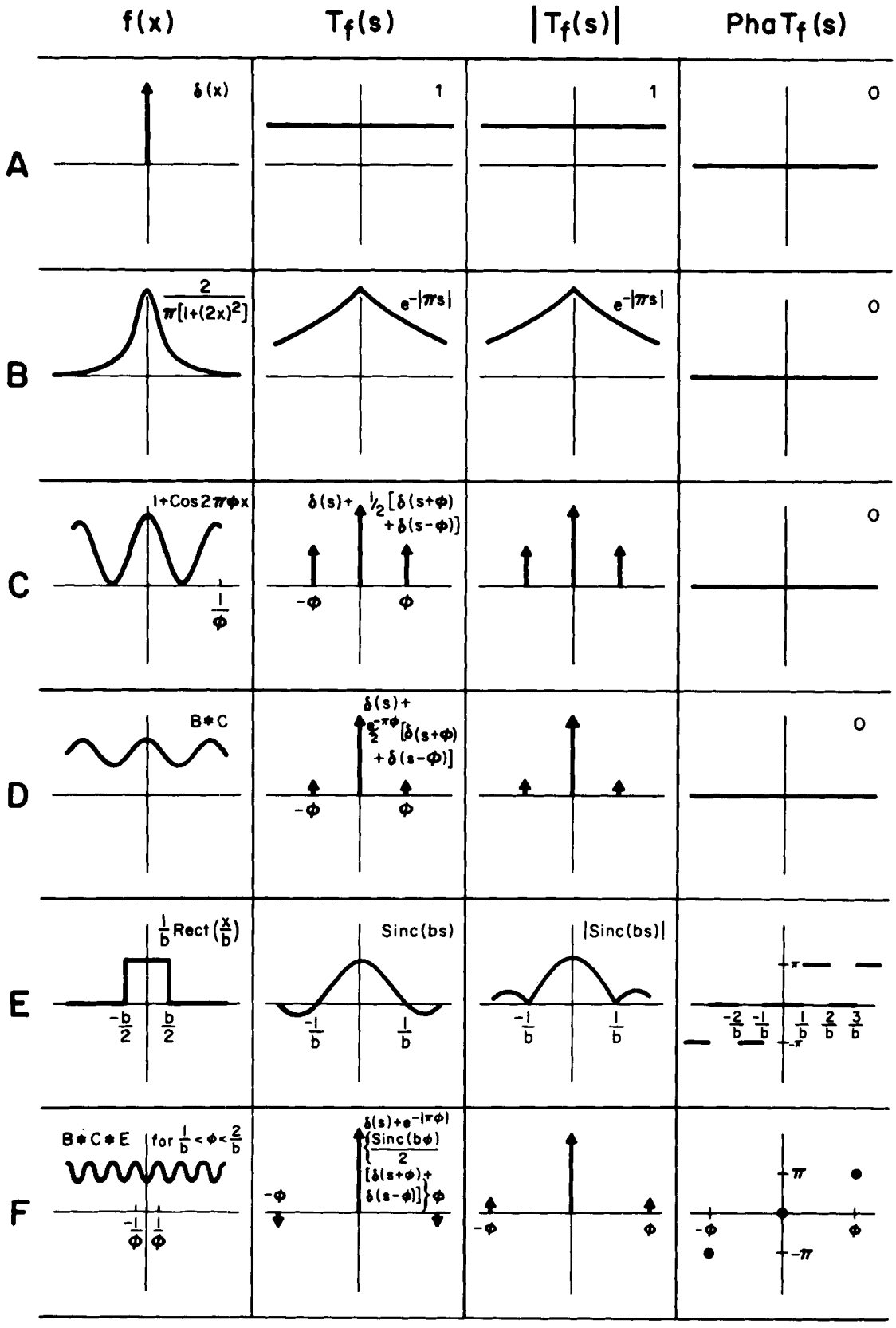

FIG. A1. Fourier analytic concepts involved in one-dimensional spurious resolution. See text for description.

earlier in Fig. 8], of frequency $\sqrt{u^2 + v^2}$, which has been rotated away from the vertical by an angle of $\tan^{-1}(v/u)$. Thus when $v = 0$ the grating is vertical; when $u = 0$ the grating is horizontal. In general, the grating $\cos[2\pi(ux + vy)]$ is perpendicular to the straight line $y = (v/u)x$. The phase term indicates how far the grating must be shifted along its axis [i.e., along the line $y = (v/u)x$]: a distance $(2\pi)^{-1}(u^2 + v^2)^{-1/2}$ pha $t_f(u,v)$ to the left. Altogether then, Equation A12 expresses the fact that $f(x,y)$ is a weighted sum of cosinusoidal gratings of various frequencies and orientations; $|t_f(u,v)|$ indicates how much to weight the cosine of frequency $\sqrt{u^2 + v^2}$ at orientation $\tan^{-1}(v/u)$, and pha $t_f(u,v)$ indicates how much to shift it, in order for the sum to reproduce f.

A two-dimensional SIL operator O is a device that transforms input functions $f(x,y)$ into output functions $O[f(x,y)]$ and satisfies the two conditions of *1*) linearity: $O[f + g] = O[f] + O[g]$; and *2*) shift invariance: if $O(x,y)$ is the output to $f(x,y)$, $O(x - a, y - b)$ is the output to $f(x - a, y - b)$ [i.e., shifting the input bodily to a new origin (a,b) shifts the output bodily to the same origin]. Such an operator can always be represented as a two-dimensional convolution

$$O[f(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x - x', y - y') \cdot o(x',y')\,dx'dy' = f(x,y)*o(x,y) \quad \text{(A13)}$$

where the function o is the impulse response of O. If the input is a two-dimensional pulse of unit energy [e.g., a very tall thin cylinder of unit volume] at the origin, the output will be exactly $o(x,y)$. In optical applications o is called the *pointspread function.* The Fourier transform of o, $t_o(u,v)$, is the transfer function of O. Knowledge of either the impulse response or the transfer function completely characterizes the operator O just as in the one-dimensional case, since the response to any input, f, can be determined either directly by the convolution $f*o$ or indirectly (i.e., in the transform domain) via a multiplicative relationship corresponding to Equation A7 in the one-dimensional case

$$t_{O[f]}(u,v) = t_f(u,v)t_o(u,v) \quad \text{(A14)}$$

Taking absolute values on both sides of Equation A14 yields the two-dimensional version of Equation A8

$$|t_{O[f]}(u,v)| = |t_f(u,v)|\,|t_o(u,v)| \quad \text{(A15)}$$

The nonnegative function $|t_o(u,v)|$ is the MTF of operator O.

Relationship Between Linespread and Pointspread Functions

Suppose the linespread function of an optical system (e.g., the eye) is $o_1(x)$ [e.g., $o_1(x) = 2\pi^{-1}[1 + 2x)^2]^{-1}$, the linespread function for the eye illustrated in Fig. A1, Sequence B]. If the system is circularly symmetric, so that rotating the object line simply rotates the original image with no change in spread, the pointspread function $o(x,y)$ (and consequently the two-dimensional transfer function) can be immediately determined. At any point in the u,v plane the two-dimensional transfer function $t_o(u,v)$ depends only on the distance from the origin $\sqrt{u^2 + v^2}$, and its value will be the one-dimensional transfer function at that distance, i.e., $t_o(u,v) = t_{o_1}(\sqrt{u^2 + v^2})$. Consequently the two-dimensional impulse response (i.e., the point spread) can be obtained by Fourier inversion

$$o(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i2\pi(ux+vy)} t_{o_1}(\sqrt{u^2 + v^2})\,du dv \quad \text{(A16)}$$

In the case of the linespread function in our illustration this yields the circularly symmetric pointspread function

$$o(x,y) = 2\pi^2(\pi^2 + 4\pi^2 r^2)^{-3/2} \quad \text{(A17)}$$

where $r = \sqrt{x^2 + y^2}$.

To determine the effect of defocusing we assume that each point $f(x,y)$ in the object is first convolved with a disk function of the form $D(x,y) = 4(\pi b^2)^{-1}$ for $\sqrt{x^2 + y^2} \leq b/2$; $D(x,y) = 0$ elsewhere [this corresponds to a blur circle of diameter b; the factor $4(\pi b^2)^{-1}$ makes the area equal one], and the result is then convolved with the pointspread function given by Equation A17. The transform of $D(x,y)$ is

$$t_o(u,v) = \left(\frac{2}{\pi}\right)\left[\frac{J_1(\pi bq)}{bq}\right] \quad \text{(A18)}$$

where $q = \sqrt{u^2 + v^2}$, and J_1 is a Bessel function of the first kind. The graph of Equation A18 (for the case $b = 2$) is illustrated in Figure A2. Notice that this two-dimensional function has negative regions, analogous to the negative regions of its one-dimensional analog sinc (bs) (Fig. A1, Sequence E). (However, in this case the zeros are not evenly spaced along the radius.) Consequently the transfer function of a misfocused eye (i.e., the product of Eqs. A18 and A17) is also negative in the same regions of the u,v plane. This means that sinusoidal inputs with spatial frequencies in these regions will appear in the image with reversed sign, i.e., spurious resolution.

Coherent and Incoherent Illumination

The discussion so far has been carried on in terms of the "intensity" of optical images—implicitly, a quantity proportional to quanta/s per unit area incident on the image plane, or emitted from the object. We have identified an input function, $f(x,y)$, with the intensity of light in the object plane at (x,y), and an output function, $O[f(x,y)]$, with intensity at (x,y) in the image plane. This description does not explicitly

acknowledge the wave properties of light, and so one might wonder how (or whether) these are taken into account. The story, very briefly, is as follows. Fourier optics is based on scalar diffraction theory (51) in which the (monochromatic) light at a spatial point (x,y) in some plane is represented by a pair of numbers $a(x,y)$, $p(x,y)$ that correspond respectively to the amplitude and phase of a sinusoidally varying electric vector at (x,y). The expression for this vector is $a(x,y)\cos[2\pi\phi t + p(x,y)]$, where t is time and ϕ is the frequency of the monochromatic light. Thus the light at (x,y) can be expressed as the real part of a complex number of the form $A(x,y) = a(x,y)e^{i[p(x,y)+2\pi\phi t]}$. The intensity of the light at (x,y) is then defined to be $|A(x,y)|^2$ [i.e., $|a(x,y)|^2$]. This quantity is proportional to quanta/s at (x,y), and consequently determines the response of any photoreceptive device (e.g., photopigments, film) in the image plane. The phase term $p(x,y)$ plays a role when the light at image point (x,y) is the sum of two or more sinusoidally varying components that may reinforce or cancel one another depending on their phase relationships. This is how the scalar diffraction model encompasses interference effects.

Now there are two cases to consider, depending on whether the light emitted from the object is coherent or incoherent. The coherent case assumes that the phases of the electric vectors at each point in the object are locked together, so that the phases $p(x,y)$ and $p(x',y')$ corresponding to any pair of object points have a difference that remains constant over time. (By definition, if there is only one object point, i.e., a point source, the light is necessarily coherent.) This can be achieved in practice by laser illumination, or (much less efficiently) by ensuring that all of the light stems from a single point source. In the incoherent case, which is the ordinary state of affairs under natural viewing conditions, the phases corresponding to different object points vary randomly and independently over time. In the image plane the coherent case allows for the possibility of stable phase relationships between contributions to $A(x,y)$ coming from many different object points, and so interference must be taken

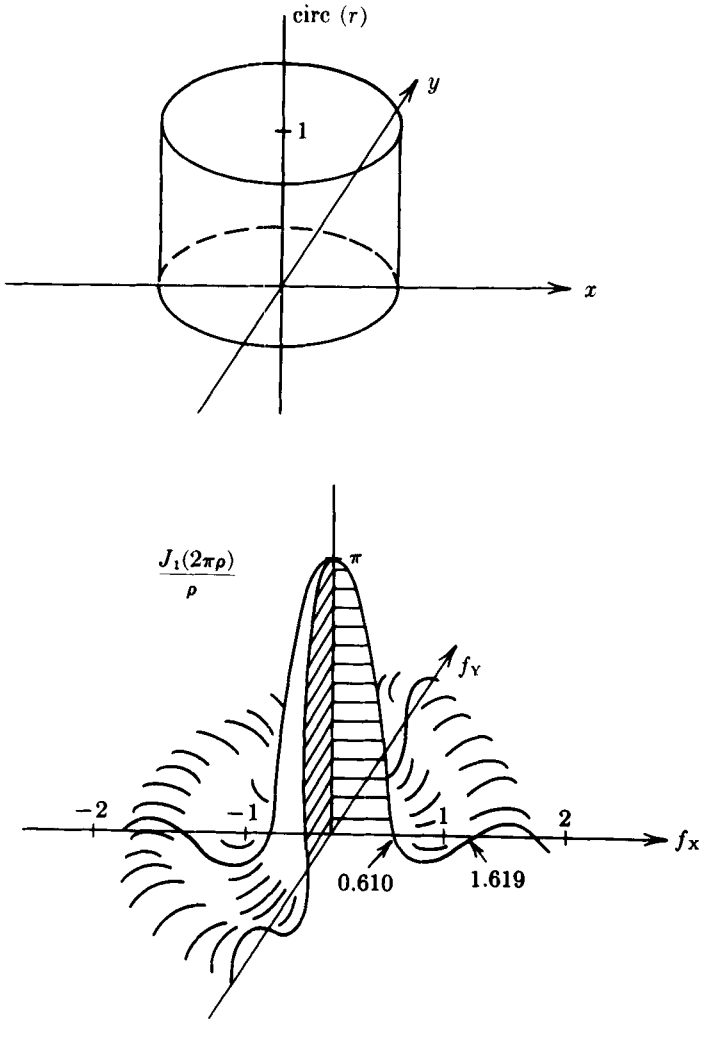

FIG. A2. Two-dimensional disk function (*top*) and its Fourier transform (*bottom*). Up to a constant factor π, the latter is the graph of Equation A18. [From Goodman (51).]

into account. In the incoherent case, on the other hand, no stable phase relationships exist between contributions stemming from different object points, and so no stable interference patterns can arise.

The two cases of coherent and incoherent illumination give rise to two different (but simply interrelated) Fourier analytic treatments. Let $A_1(x,y)$ denote the complex electric vector $a(x,y)e^{i[p(x,y)+2\pi\phi t]}$ at point (x,y) in the object plane, and $A_2(x,y)$ the electric vector at (x,y) in the image. [It is assumed that the x and y axes are parametrized so that (x,y) in the image plane is the geometrical optics image of (x,y) in the object plane, i.e., by measuring both in terms of visual angle.] And let $M(x,y)$ denote the complex impulse response produced by an object consisting of a single point source at the object point (0,0). Such an object is necessarily coherent, and $M(x,y)$ is the complex response $|M(x,y)|e^{i[2\pi\phi t+\mathrm{pha}M(x,y)]}$ representing both the amplitude and phase of the electric vector at image point (x,y) produced by a point object at (0,0). Now consider an extended object consisting of many points. The quantity of interest is $|A_2(x,y)|^2$, the intensity at image point (x,y), since this determines quantum density on the photoreceptive surface in the image plane. Then it can be shown (17) that the coherent object case gives rise to the relationship

$$|A_2(x,y)|^2 = |M(x,y)*A_1(x,y)|^2 \qquad \text{(A19)}$$

[where, as usual, $(*)$ denotes convolution], while the incoherent case gives rise to

$$|A_2(x,y)|^2 = |M(x,y)|^2*|A_1(x,y)|^2 \qquad \text{(A20)}$$

The second case is the one assumed in all of the earlier discussions. It expresses the fact that the intensity of the image (i.e., $|A_2|^2$) is the convolution of the intensity of the object $|A_1|^2$ with the intensity impulse response, $|M(x,y)|^2$. In terms of the notation used earlier, $|A_1|^2$ is the input function $f(x,y)$, $|M|^2$ is the impulse response, $o(x,y)$, and $|A_2|^2$ is the output function $\circ[f(x,y)]$. So Equation A20 corresponds exactly to the earlier relationship Equation A13. Equation A19, on the other hand, describes the result expected with coherent objects, which is quite different, since in general $|M * A|^2$ is not the same as $|M|^2 * |A|^2$. For example, M may be negative, but $|M|^2$ can never be. It is interesting to note that when the object is a single point (i.e., necessarily coherent) so that the appropriate expression for $|A_2(x,y)|^2$ is Equation A19, the result is $|A_2(x,y)|^2 = |M(x,y)|^2$, which is the impulse response for the incoherent case. Consequently the coherent impulse response $M(x,y)$ is never observed in isolation, even with coherent illumination.

Sampling Theorem

Suppose $f(x)$ is a function (e.g., a light-intensity profile) which we can only observe at a series of equally spaced discrete values on the x axis, i.e., we can only know $f(0)$, $f(w)$, $f(-w)$, $f(2w)$, $f(-2w)$, ..., where w is the distance between successive sample points. (This is essentially the way a one-dimensional array of equally spaced thin photoreceptors views the retinal image. A model for receptors that have appreciable width relative to the spatial periods of the input is given below.) How does such a sampling affect our ability to reconstruct the entire function f? Fourier analysis provides an elegant answer to this question in the form of the *sampling theorem*: If the Fourier transform of f vanishes for frequencies greater than some cutoff frequency (s_c) [i.e., $t_f(s) = 0$ for $|s| > s_c$], $f(x)$ can be perfectly reconstructed from a series of equally spaced sample values; $f(0)$, $f(w)$, $f(-w)$, ..., $f(nw)$, $f(-nw)$, ..., provided the sampling rate, w^{-1}, is greater than or equal to $2(s_c)$. The proof of the sampling theorem is straightforward but a bit too long to include here (see ref. 27). However, Figure A3 (from ref. 27) illustrates the underlying idea and also shows what happens when the sampling rate is too slow (i.e., less than twice the highest frequency in the input). In the frequency domain, the effect of sampling $f(x)$ at discrete intervals is to create a new function whose transform is the sum of a series of replicas of the original transform t_f. [Note: the figure uses $F(s)$ to denote the transform of $f(x)$.] These replicas are centered at 0, $1/w$, $-1/w$, ... n/w, $-n/w$, ... When the sampling rate $1/w$ is $\geq 2s_c$, these replicas do not overlap on the s axis. Consequently the replica centered at the origin provides an undistorted image of the original transform t_f, which can be used to reconstruct an undistorted image of the original function f. However, if the sampling rate is less than $2s_c$, so that the replicas on the s axis overlap, the replicas centered at higher frequencies add nonzero values to the one centered at the origin, so that it is no longer an exact image of t_f, and consequently the reconstruction process will yield a distorted version of f. This kind of distortion is known as "aliasing," because its effect is that high-frequency components in the input f reappear in the sampled version of f in the guise of low-frequency components. This can have the effect of introducing frequency components in the sampled version which were nonexistent in the original function. Consequently, in order for a sampling scheme to yield undistorted reconstructions of all potential inputs, it is essential to set the sampling rate equal to (or greater than) twice the highest frequency that will occur in any input. (It does not matter if the rate is higher than needed for some inputs, since sampling too fast does not introduce any distortion.) Figure A4 illustrates the aliasing effect.

In two dimensions, sampling consists of observing $f(x,y)$ at a set of discrete points in the x,y plane (e.g., a square lattice of the form (wn,wm), $n = 0, 1, \ldots$, $m = 0, 1, \ldots$). In this case the optimal sampling scheme depends on the shape of the region in the u,v plane for which $t_f(u,v)$ is nonzero. The basic idea is still that f

can be exactly reconstructed from a discrete set of sample values, provided t_f vanishes outside some region, but the appropriate spacing of these samples depends on the shape of that region, and consequently the two-dimensional sampling theorem cannot be stated quite as neatly as the one-dimensional version.

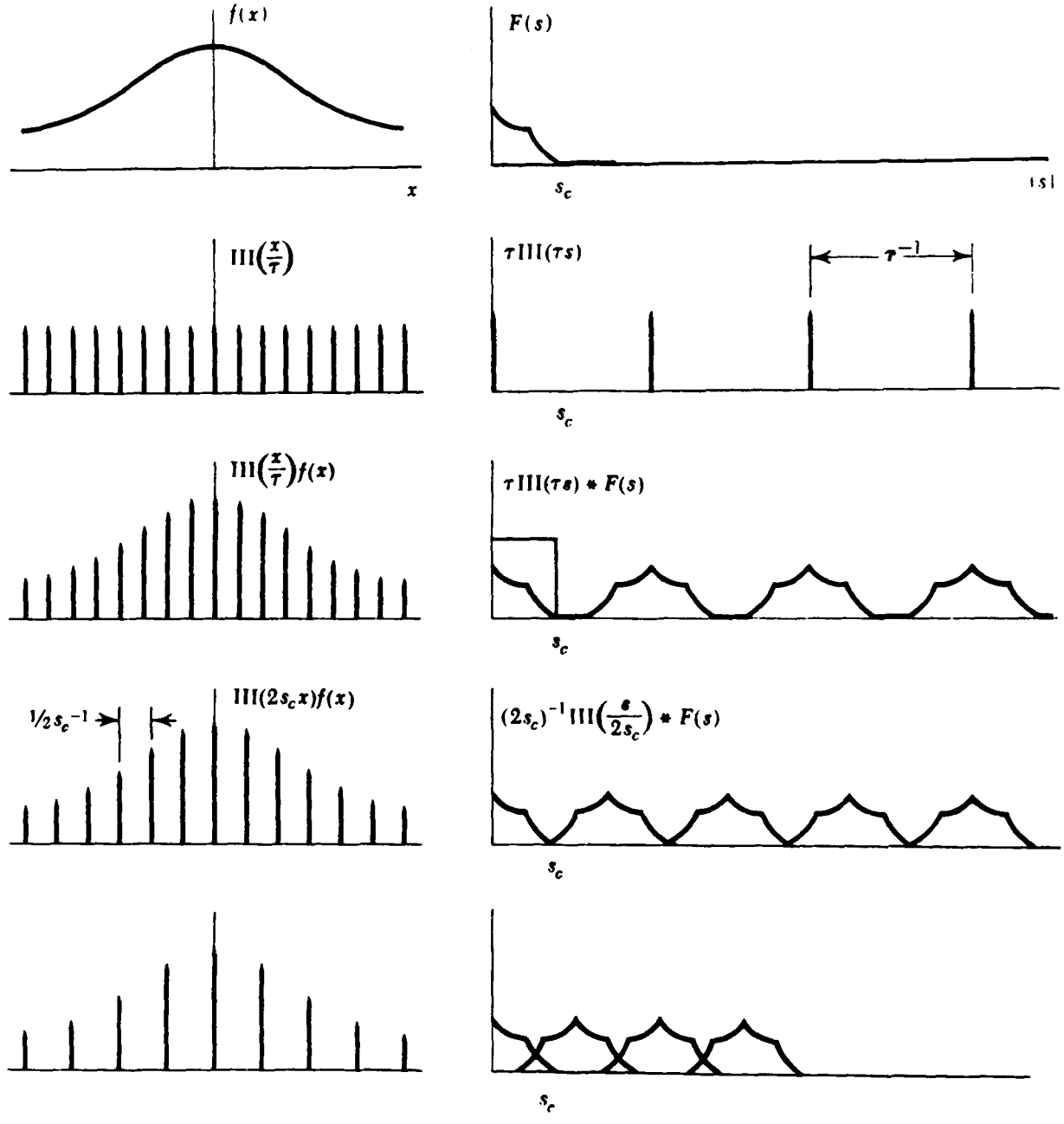

FIG. A3. Concepts involved in the one-dimensional sampling theorem; $f(x)$ is an input signal with transform $F(s)$ which vanishes for $s > s_c$. III (x/t) is an infinite row of delta functions (i.e., sample points) spaced t units apart; its transform is a row of delta functions spaced t^{-1} units apart. The *third row* shows the sampled version of f (i.e., the product $f(x)$ III (x/t)) and its transform. Note that here the sampling rate is inefficiently high: successive replicas of $F(s)$ are separated by empty intervals. The *fourth row* illustrates the optimal sampling scheme, where $t = (2s_c)^{-1}$: Here the replicas of $F(s)$ in the transform of the sampled signal are precisely adjacent. The *fifth row* shows the effect of sampling too coarsely: In the transform of the sampled signal the replicas of $F(s)$ overlap (aliasing). [From Bracewell (27). *The Fourier Transform and Its Applications* (2nd ed.), by R. Bracewell. Copyright © 1978, McGraw-Hill Book Company. Used with permission of McGraw-Hill Book Company.]

Probably the most useful exact statement can be based on the assumption that the two-dimensional transform of $f(x,y)$, i.e., $t_f(u,v)$, vanishes outside some square region in the u,v plane, i.e., $u \leq c$, $v \leq c$ [that is, the smallest square enclosing the actual nonzero region of $t_f(u,v)$]. In this case f can be exactly reconstructed by a square lattice of sample points $x = wn$, $y = wm$, with $w \leq (2c)^{-1}$ (see ref. 17).

Application to Sampling by Photoreceptors

Consider a one-dimensional retina consisting of a tightly packed line of photoreceptors, each with diameter w (e.g., an array of cones along the x axis of the retina). Suppose the input to this model retina is a function $f(x)$ (e.g., a light-intensity profile along the x axis), and that each receptor integrates the portion of f which lies directly above it and outputs that value. Thus, for example, the output of the receptor centered at the origin is $\int_{-w2/2}^{w/2} f(x)\,dx$. How much information about the input function f can be obtained from the output of the photoreceptor array? Notice that this model is not exactly the same as the case envisioned by the sampling theorem, because our receptors do not report the values of $f(x)$ at $x = 0, \pm w \pm 2w$, but rather the values of $\int f(x)dx$ over intervals of width w centered at $x = 0, \pm w, \pm 2w, \ldots$. However this simply means that we are sampling the function $F(x + w/2) - F(x - w/2)$, where $F(x) = \int_{-\infty}^{x} f(y)\,dy$, and it can readily be shown that the resulting output function takes the form

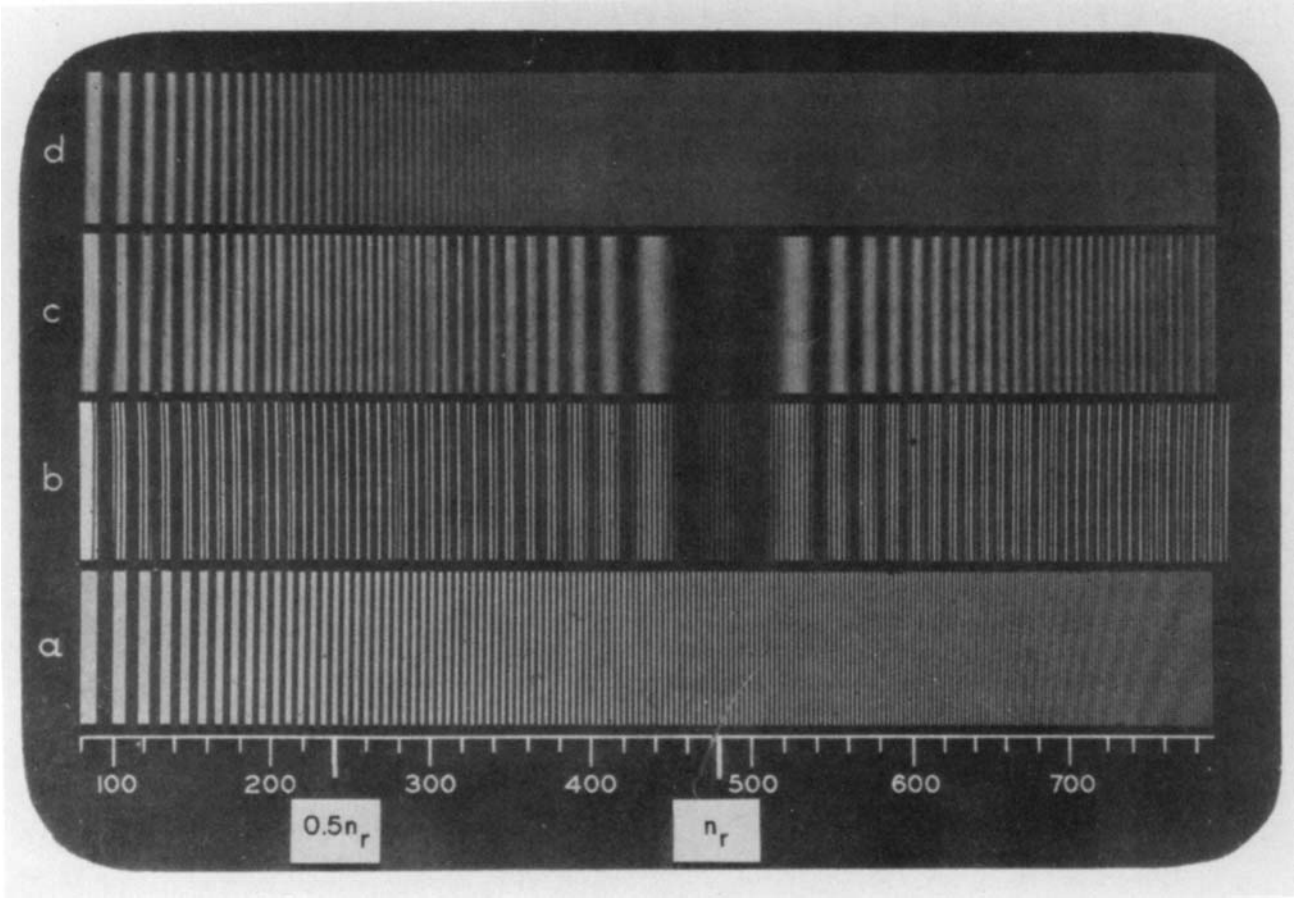

FIG. A4. One-dimensional aliasing effects. *Row a* is a linearly increasing frequency pattern. *Row b* shows the sampled appearance of pattern *a* viewed through a raster plate consisting of fine vertical slits spaced at a frequency equal to n_r on the scale. (This sampling rate guarantees perfect reconstruction for frequencies up to 0.5 n_r, with aliasing distortion for higher frequencies.) *Row c* shows the effect of spatially postfiltering the sampled image *b* (in this case by adjusting the scanning spot size in a television camera which views pattern *b*. The visual analogue would be lateral neural interactions at the receptors or beyond.) Notice that this removes the raster lines but does not eliminate aliasing: Frequencies higher than 0.5 n_r still appear as lower frequencies. *Row d* shows the effect of the spatial postfilter alone (i.e., on pattern *a* viewed without the raster plate). [From Schade (116).]

output to input $f(x)$

$$= \frac{1}{w}\,\mathrm{comb}\left(\frac{x}{w}\right)\cdot\left[f(x) * \mathrm{rect}\left(\frac{x}{w}\right)\right] \tag{A21}$$

where $1/w$ comb(x/w) denotes a series of impulses of unit energy (i.e., a series of delta functions as in Fig. 34) spaced at 0, $\pm w$, $\pm 2w$, ..., and rect(x/w) is a function which equals one for $-w/2 \le x \le w/2$ and zero elsewhere ($*$ denotes convolution). Now the factor $1/w$ comb$\left(\frac{x}{w}\right)$ represents a sampling scheme with sampling rate w^{-1}, and Equation A21 means that this scheme is applied to the convolution of the input f with a rectangular function of width w. The Fourier transform of this convolution is $t_f(s)\cdot w$ sinc(ws) [where sinc$(y) = (\sin \pi y)/\pi y$ is the function depicted in Fig. 32], and the transform of the receptor output (i.e., of Eq. A21) is

$$w\,\mathrm{comb}(ws) * [\mathrm{sinc}(ws)t_f(s)] \tag{A22}$$

Thus Equation A21 tells us that the receptor output represents a sampling at rate $1/w$ of a function whose transform is essentially the product $t_f(s)$ sinc(ws), which has zeros at $s = \pm 1/w, \pm 2/w \ldots$ (because of the sinc factor). Beyond the first zero sinc(ws) becomes negative (for a time), and so beyond this point the reproduction scheme introduces spurious resolution. Consequently we concentrate on input frequencies $<1/w$—i.e., consider the effect of the scheme on inputs with frequencies below the sampling rate. According to the sampling theorem, the sampling scheme $1/w$ comb(x/w) is perfect for inputs with frequencies $s < (2w)^{-1}$, and one can conclude that sinusoidal input in this range can be recovered with no frequency distortion, though the amplitude of each input frequency s will be attenuated by the factor sinc(ws). For input frequencies in the range $(2w)^{-1} < s < w^{-1}$, however, sampling by receptors of width w leads to aliasing distortion in the output.

To see explicitly the form this aliasing takes, consider a cosinusoidal input with frequency ϕ, with $(2w)^{-1} < \phi < w^{-1}$. In this case the spectrum of the output of receptor sampling will contain energy at $s = w^{-1} - \phi$ (which lies in the interval $[0, (2w)^{-1}]$), and also at $s = w^{-1} + \phi$, plus all other points of the form $s = nw^{-1} \pm \phi$, $n = 0, \pm 1, \pm 2, \ldots$. This output spectrum is identical to the one that is produced by an input cosine of frequency $w^{-1} - \phi$, except that each spike in the output spectrum of $\cos[2\pi\phi x]$ is attenuated (relative to the output spectrum of $\cos[2\pi(w^{-1} - \phi)x]$) by the constant factor $\mathrm{sinc}(w\phi)/\mathrm{sinc}(1 - w\phi) = r(\phi)$. Consequently, the sampled output of input $\cos 2\pi\phi x$ is indistinguishable from the sampled output produced by inputting the lower frequency function $r(\phi)\cos[2\pi(w^{-1} - \phi)x]$, and any reconstruction process must confuse these two inputs.

Finally, it is natural to wonder about the effect of shrinking receptor width while keeping the spacing between receptor centers the same, to model the effect of gaps between receptors. Specifically, suppose the width of each receptor is reduced from w to pw, $0 \le p \le 1$, while the spacing between receptor centers is kept at w. In this case the output is

$$\frac{1}{w}\,\mathrm{comb}\left(\frac{x}{w}\right)\left[f(x) * \mathrm{rect}\left(\frac{x}{pw}\right)\right] \tag{A23}$$

which represents a sampling at rate w^{-1} of a function $\left[\text{i.e., } f(x) * \mathrm{rect}\left(\frac{x}{pw}\right)\right]$ which has transform, $t_f(s)\cdot pw\,\mathrm{sinc}(pws)$. Consequently the effect of shrinking receptor width is not to increase the effective frequency bandwidth of the system (i.e., for transmission with no harmonic distortion), since aliasing will still occur for $s > (2w)^{-1}$. However, it will increase the

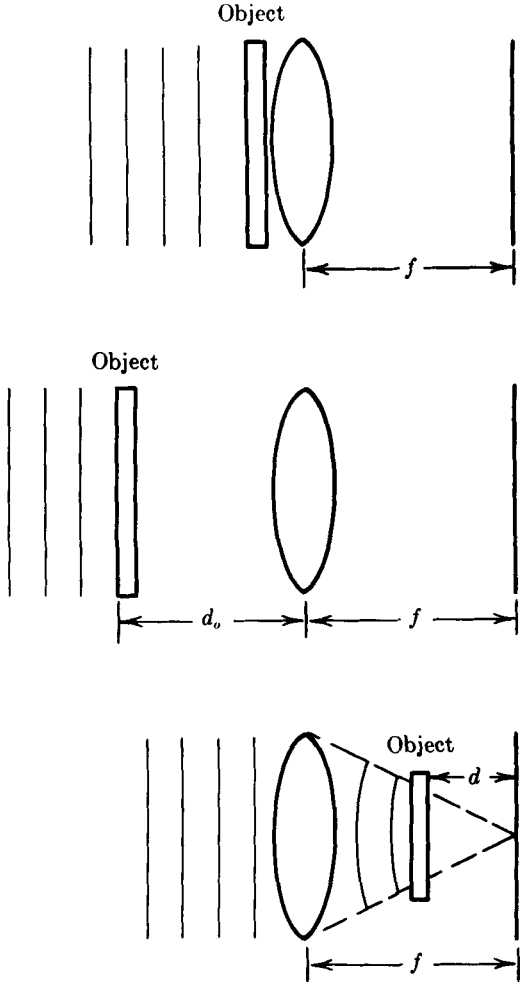

FIG. A5. Configurations for producing optical Fourier transforms. The focal length of the lens is represented by f; thin vertical lines represent the illumination (a monochromatic plane wave). [From Goodman (51).]

range between the cutoff frequency $(2w)^{-1}$ and the point (i.e., $s = (pw)^{-1}$) at which spurious resolution begins to occur.

The same ideas can be carried over to two dimensions by imagining (for the sake of convenience) square receptors with areas $(pw)^2$ centered at the points $x = 0, \pm w, \pm 2w, \ldots; y = 0, \pm w, \pm 2w, \ldots$. If each receptor integrates the portion of $f(x,y)$ directly above it, the resulting output is

$$w^{-2}\,\mathrm{comb}(x/w)\,\mathrm{comb}(y/w)$$
$$\cdot[f(x,y) * \mathrm{rect}(x/pw)\,\mathrm{rect}(y/pw)]$$

The factor outside the square brackets is a square lattice sampling scheme that provides reproduction without harmonic distortion for all two-dimensional inputs whose transforms vanish outside the square $u \leq (2w)^{-1}$, $v \leq (2w)^{-1}$. Consequently, if the highest spatial frequency in the input at any orientation is c, the required receptor spacing w is $(2c)^{-1}$, just as in the one-dimensional case. Any sparser sampling will lead to aliasing.

Optical Transforms and Spatial Filtering

Calculating the two-dimensional Fourier transforms of input functions corresponding to complex natural scenes (e.g., photographs of real objects) is extremely laborious, but for many purposes the same result can be obtained directly by elegant optical methods requiring no calculation at all. These methods depend on the fact that the diffraction pattern formed by an aperture is essentially the Fourier transform of that aperture (51). Three optical configurations used to compute transforms are illustrated in Figure A5. In each case the "object" is a screen whose transparency can be described by a transmission function $T(x,y)$ (which varies between zero and one) and the illumination (represented by the vertical lines to the left of the object) is assumed to be a coherent monochro-

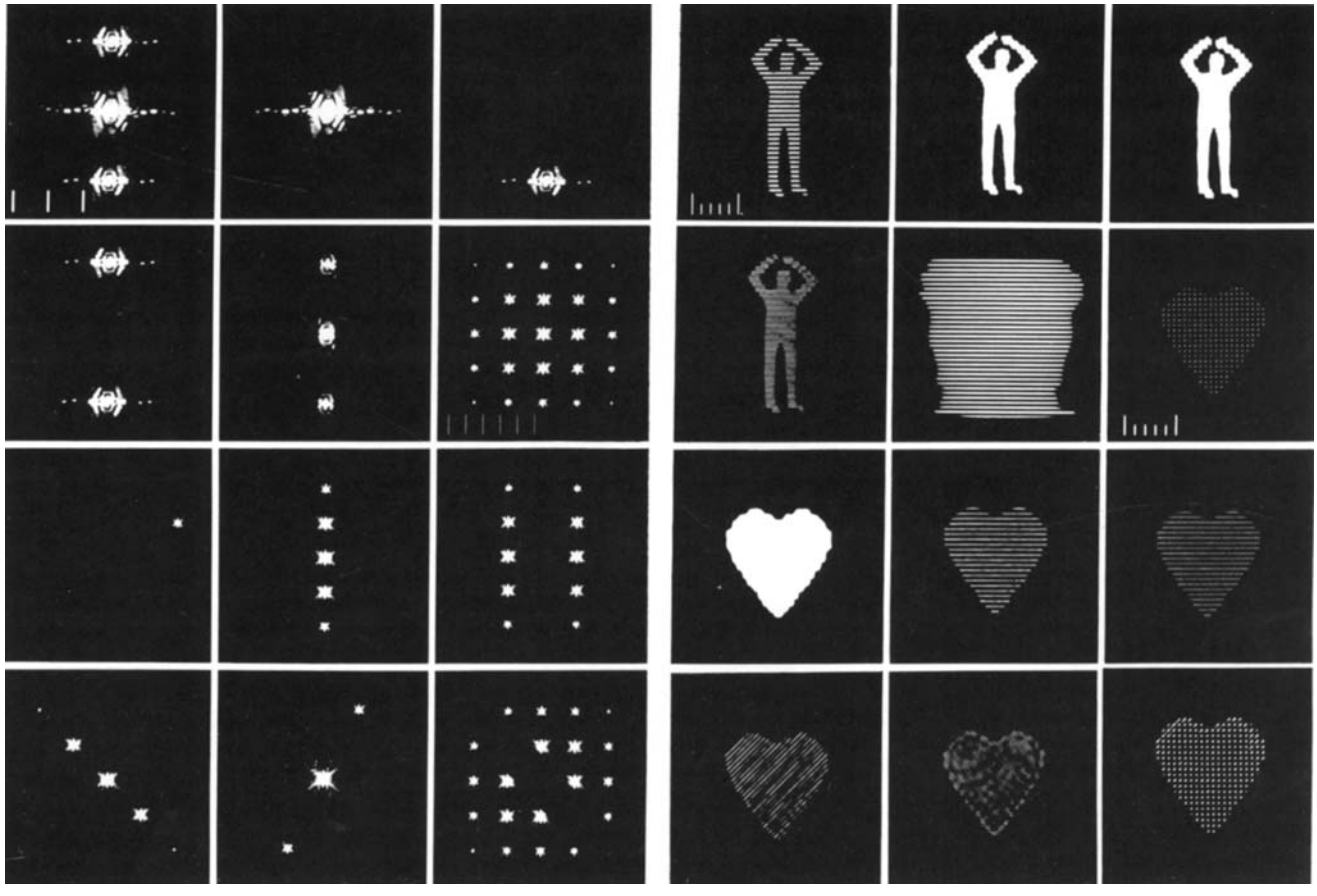

FIG. A6. Optical Fourier transforms. Each *panel* in the *left block* (three *left-hand columns*) shows the optical transform (i.e., amplitude spectrum) of the corresponding image in the *right block* (three *right-hand columns*). [From Harburn et al. (61). Reprinted from G. Harburn, C. A. Taylor, and T. R. Welberry: *Atlas of Optical Transforms*. Copyright © 1975 by G. Bell and Sons, Ltd. Used by permission of the publisher, Cornell University Press.]

matic plane wave (e.g., from a laser). All three configurations will result in an image on the screen at the right whose intensity distribution $I(x,y)$ is essentially the squared absolute value of the Fourier transform of $T(x,y)$. (That is, the image intensity will be proportional to the squared amplitude spectrum of the transmission function of the object, the proportionality constant depending on the intensity and wavelength of the light and the focal length of the lens.) A number of complex objects and their optical transforms (in effect, their amplitude spectra) are illustrated in Figure A6. The figure also shows the effects of spatial filtering, which can readily be achieved by placing a stop in the transform plane to block certain spatial frequencies, and then retransforming the image in the transform plane (i.e., by repeating the optical transform process using the image in the transform plane as an object). The result is a filtered version of the original object, i.e., $T(x,y)$ minus the blocked spatial frequencies.

The same principle naturally applies to the eye, the optics of which can serve as the lens in Figure A5. This is the basis of the technique employed by Le Grand and others for circumventing the normal spatial frequency cutoff of the eye, as discussed in VISUAL ACUITY, p. 260.

We thank R. Baillargeon, N. Graham, M. Hayhoe, and D. MacLeod for helpful suggestions. Financial support from the U.S. National Institutes of Health (Grant 1-R01-EY03164-01 to B. Wandell) and the University of California, Irvine (Focused Research Project on Perception and Higher Mental Processes) is gratefully acknowledged.

REFERENCES

1. ABNEY, W., AND E. R. FESTING. Colour photometry. *Philos. Trans. R. Soc. London* 177: 423–456, 1886.
1a. AGUILAR, M., AND W. S. STILES. Saturation of the rod mechanism of the retina at high levels of stimulation. *Opt. Acta* 1: 59–64, 1954.
2. ALPERN, M. Rod vision. In: *The Assessment of Visual Function,* edited by A. M. Potts. St. Louis: Mosby, 1972.
3. ALPERN, M., AND J. MOELLER. The red and green cone visual pigments of deuteranomalous trichromacy. *J. Physiol. London* 266: 647–675, 1977.
4. ALPERN, M., AND E. N. PUGH, JR. Variation in the action spectrum of erythrolable among deuteranopes. *J. Physiol. London* 266: 613–646, 1977.
5. ALPERN, M., AND T. WAKE. Cone pigments in human deutan colour vision defects. *J. Physiol. London* 266: 595–612, 1977.
6. ARDEN, G. B., AND R. A. WEALE. Nervous mechanisms and dark-adaptation. *J. Physiol. London* 125: 417–426, 1954.
7. BARLOW, H. B. Increment threshold at low intensities considered as signal-noise discriminations. *J. Physiol. London* 119: 69–88, 1957.
8. BARLOW, H. B. Temporal and spatial summation in human vision at different background intensities. *J. Physiol. London* 141: 337–350, 1958.
9. BARLOW, H. B. Dark-adaptation: a new hypothesis. *Vision Res.* 4: 47–58, 1964.
10. BARLOW, H. B. Dark and light adaptation: psychophysics. In: *Handbook of Sensory Physiology. Visual Psychophysics,* edited by L. Hurvich and D. Jameson. Berlin: Springer-Verlag, 1972, vol. 7, pt. 4, p. 1–28.
10a. BARLOW, H. B. Retinal and central factors in human vision limited by noise. In: *Vertebrate Photoreception,* edited by H. B. Barlow and P. Fatt. New York: Academic, 1977, p. 337–358.
11. BARLOW, H. B., C. BLAKEMORE, AND J. D. PETTIGREW. The neural mechanism of binocular depth discrimination. *J. Physiol. London* 193: 327–342, 1967.
12. BARLOW, H. B., AND P. FATT (editors). *Vertebrate Photoreception.* New York: Academic, 1977.
13. BARLOW, H. B., AND B. SAKITT. Doubts about scotopic interactions in stabilized vision. *Vision Res.* 13: 523–524, 1973.
14. BARLOW, H. B., AND J. M. B. SPARROCK. The role of afterimages in dark adaptation. *Science* 144: 1309–1314, 1964.
16. BAYLOR, D. A., AND R. FETTIPLACE. Transmission from photoreceptors to ganglion cells in the retina of the turtle. In: *Vertebrate Photoreceptors,* edited by H. B. Barlow and P. Fatt. New York: Academic, 1977, p. 193–203.
17. BAYLOR, D. A., M. G. F. FUORTES, AND P. M. O'BRYAN. Receptive fields of cones in the retina of the turtle. *J. Physiol. London* 214: 265–294, 1971.
18. BAYLOR, D. A., AND A. L. HODGKIN. Changes in time scale and sensitivity in turtle photoreceptors. *J. Physiol. London* 242: 729–758, 1974.
19. BEDFORD, R. E., AND G. WYSZECKI. Axial chromatic aberration of the eye. *J. Opt. Soc. Am.* 47: 564–565, 1957.
20. BERKLEY, M. A., F. KITTERLE, AND D. W. WATKINS. Grating visibility as a function of orientation and retinal eccentricity. *Vision Res.,* 15: 239–244, 1975.
21. BLAKEMORE, C., AND F. W. CAMPBELL. Adaptation to spatial stimuli. *J. Physiol. London* 200: 11–13, 1969.
22. BLAKEMORE, C. B., AND W. A. H. RUSHTON. Dark adaptation and increment threshold in a rod monochromat. *J. Physiol. London* 181: 612–628, 1965.
23. BLAKEMORE, C. B., AND W. A. H. RUSHTON. The rod increment threshold during dark adaptation in normal and rod monochromat. *J. Physiol. London* 181: 629–640, 1965.
24. BLICK, D. W., AND D. I. A. MACLEOD. Rod threshold: influence of neighboring cones. *Vision Res.* 18: 1611–1616, 1978.
25. BOYNTON, R. M. Ten years of research with the minimally distinct border. In: *Visual Psychophysics and Physiology,* edited by J. C. Armington, J. Krauskopf, and B. R. Wooten, New York, Academic, 1978.
26. BOYNTON, R. M., AND W. S. BARON. Sinusoidal flicker characteristics of primate cones in response to heterochromatic stimuli. *J. Opt. Soc. Am.* 65: 1091–1100, 1975.
27. BRACEWELL, R. *The Fourier Transform and its Applications* (2nd ed.). New York: McGraw-Hill, 1978.
28. BREITMEYER, B. G., AND L. GANZ. Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychol. Rev.* 83: 1–36, 1976.
29. BRINDLEY, G. S. *Physiology of the Retina and Visual Pathway.* Baltimore, Williams and Wilkins, 1970.
29a. BRINK, G. VAN DEN. Measurements of the geometrical aberrations of the eye. *Vision Res.* 2: 233–244, 1962.
30. BROWN, J. L., M. P. KUHNS, AND H. E. ADLER. Relation of threshold criterion to the functional receptors of the eye. *J. Opt. Soc. Am.* 47: 198–204, 1957.
31. BYRAM, G. M. The physical and photochemical basis of visual resolving power. Part III. Visual acuity and the photochemistry of the retina. *J. Opt. Soc. Am.* 34: 718–738, 1944.
32. CAMPBELL, F. W., AND D. G. GREEN. Optical and retinal factors affecting visual resolution. *J. Physiol. London* 181: 576–593, 1965.
33. CAMPBELL, F. W., AND R. W. GUBISCH. Optical quality of the

human eye. *J. Physiol. London* 186: 558–578, 1966.
34. CAMPBELL, F. W., AND J. G. ROBSON. Application of Fourier analysis to the visibility of gratings. *J. Physiol. London* 197: 551–566, 1968.
35. CARTERETTE, E. C., AND M. P. FRIEDMAN (editors). *Handbook of Perception. Biology of Perceptual Systems.* New York: Academic, 1973, vol. III.
35a. CARTERETTE, E. C., AND M. P. FRIEDMAN (editors). *Handbook of Perception. Seeing.* New York: Academic, 1975, vol. V.
36. CORNSWEET, T. N. *Visual Perception.* New York: Academic, 1970.
36a. COWAN, T. D. Some remarks on channel bandwidths for visual contrast detection. In: E. Poppel, R. Held, and J. E. Dowling. Neuronal mechanisms in visual perception. *Neurosci. Res. Program Bull.* 7, 15, 3. Cambridge, MA: MIT Press, 1977, p. 492–515.
37. CRAIK, K. J. W., AND M. D. VERNON. The nature of dark adaptation. *Br. J. Psychol.* 32: 64–81, 1941.
38. CRAWFORD, B. H. Visual adaptation in relation to brief conditioning stimuli. *Proc. R. Soc. London Ser. B* 134: 283–302, 1947.
39. CRAWFORD, B. H. The Stiles-Crawford effects and their significance in vision. In: *Handbook of Sensory Physiology. Visual Psychophysics*, edited by L. Hurvich and D. Jameson. Berlin: Springer-Verlag, 1972, vol. 7, pt. 4, p. 470–483.
40. DAVSON, H. (editor). *The Eye. Visual Function in Man.* New York: Academic, 1976, vol. 2A.
41. DAW, N. W., AND J. M. ENOCH. Contrast sensitivity, Westheimer function and Stiles-Crawford effect in a blue cone monochromat. *Vision Res.* 13: 1669–1680, 1973.
42. DE VALOIS, R. L., AND K. K. DE VALOIS. Neural coding of color. In: *Handbook of Perception. Seeing*, edited by E. C. Carterette and M. P. Friedman. New York: Academic, 1975, vol. V, p. 117–166.
44. DE VRIES, H. The quantum character of light and its bearing upon the threshold of vision, the differential sensitivity and acuity of the eye. *Physica* 10: 553–564, 1943.
45. DUKE-ELDER, S., AND D. ABRAMS. *Ophthalmalic Optics and Refraction.* St. Louis: Mosby, 1970.
46. ENROTH-CUGELL, C., B. G. HERTZ, AND P. LENNIE. Convergence of rod and cone signals in the cat's retina. *J. Physiol. London* 269: 297–318, 1977.
47. FAIN, G. The threshold signal of photoreceptors. In: *Vertebrate Photoreception*, edited by H. Barlow and P. Fatt. New York: Academic, 1977.
48. FLAMANT, F., AND W. S. STILES. The directional and spectral sensitivites of the retinal rods to adapting fields of different wave-lengths. *J. Physiol. London* 107: 187–202, 1948.
49. FRUMKES, T. E., AND L. A. TEMMED. Rod-cone interaction in human-scotopic vision: II. Cones influence increment thresholds detected by rods. *Vision Res.* 17: 673–679, 1977.
50. GAZZANIGA, M. S., AND C. B. BLAKEMORE (editor). *Handbook of Psychobiology.* New York: Academic, 1975.
51. GOODMAN, J. W. *Introduction to Fourier Optics.* New York: McGraw-Hill, 1965.
52. GORRAND, J. M. Diffusion of the human retina and quality of the optics of the eye on the fovea and the peripheral retina. *Vision Res.* 8: 907–912, 1979.
53. GRAHAM, N. Visual detection of aperiodic spatial stimuli by probability summation among narrowband channels. *Vision Res.* 17: 637–652, 1977.
54. GRAHAM, N., AND J. NACHMIAS. Detection of grating patterns containing two spatial frequencies: a comparison of single-channel and multiple-channels models. *Vision Res.* 11: 251–259, 1971.
55. GRAHAM, N., AND F. RATLIFF. Quantitative theories of the integrative action of the retina. In: *Contemporary Developments in Mathematical Psychology. Measurement, Psychophysics, and Neural Information Processing,* edited by D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes. San Francisco: Freeman, 1974, vol II, p. 306–371.
56. GRANIT, R. The dark adaptation of mammalian visual receptors. *Acta Physiol. Scand.* 7: 216–220, 1944.
57. GREEN, D. G. The contrast sensitivity of the colour mechanisms of the human eye. *J. Physiol. London* 196: 415–429, 1968.
58. GREEN, D. G. Regional variations in the visual acuity for interference fringes on the retina. *J. Physiol. London* 207: 351–356, 1970.
58a. GREGORY, R. L. Stereovision and isoluminence. *Proc. R. Soc. London Ser. B* 204: 467–476, 1979.
59. GUBISCH, R. W. Optical performance of the human eye. *J. Opt. Soc. Am.* 57: 407–415, 1967.
60. GUTH, S. L., N. J. DONLEY, AND R. T. MARROCCO. On luminance additivity and related topics. *Vision Res.* 9: 537–575, 1969.
61. HARBURN, G., C. A. TAYLOR, AND T. R. WELBERRY. *Atlas of Optical Transforms.* Ithaca, NY: Cornell Univ. Press, 1975.
62. HECHT, E., AND A. ZAJAC. *Optics.* Menlo Park, CA: Addison-Wesley, 1974.
63. HECHT, S. Photochemistry of visual purple. I. The kinetics of the decomposition of visual purple by light. *J. Gen. Physiol.* 3: 1–13, 1920.
64. HECHT, S. Rods, cones, and the chemical basis of vision. *Physiol. Rev.* 17: 239–290, 1937.
65. HECHT, S., C. HAIG, AND A. M. CHASE. The influence of light adaptation on subsequent dark adaptation of the eye. *J. Gen. Physiol.* 20: 831–850, 1937.
66. HELD, R., H. W. LEIBOWITZ, AND H.-L. TEUBER. *Handbook of Sensory Physiology. Perception.* Berlin: Springer-Verlag, 1978, vol. 8.
66a. HERING, E. Grundzinge des Lehre von Lichtsinn. In: *Handbuch der gesammten Augenheilkunde*, edited by A. Graefe and E. T. Saemich. Leipzig: Eugelmann, 1905, vol. 3. [Transl. L. M. Hurvich and D. Jameson. *Outlines of a Theory of the Light Sense.* Cambridge, MA: Harvard Univ. Press, 1964.]
67. HOWLAND, H. C., AND B. HOWLAND. A subjective method for the measurement of monochromatic aberrations of the eye. *J. Opt. Soc. Am.* 67: 1508–1518, 1977.
68. HUBEL, D. H., AND T. N. WIESEL. Brain mechanisms of vision. *Sci. Am.* 241: 150–162, 1979.
69. INGLING, C. R. The spectral sensitivity of the opponent-colors channel. *Vision Res.* 17: 1083–1090, 1977.
70. INGLING, C. R., AND B. H. TSOU. Orthogonal combination of three visual channels. *Vision Res.* 17: 1075–1082, 1977.
71. INGLING, C. R., B. H. P. TSOU, T. J. GAST, S. A. BURNS, J. O. EMERICK, AND L. RIESENBERG. The achromatic channel. I. The non-linearity of minimum-border and flicker matches. *Vision Res.* 18: 379–390, 1978.
72. IVANOFF, A. About the spherical aberration of the eye. *J. Opt. Soc. Am.* 46: 901–903, 1956.
73. JAMESON, D., AND L. M. HURVICH (editors). *Handbook of Sensory Physiology. Visual Psychophysics.* Berlin: Springer-Verlag, 1972, vol. 7, pt. 4.
74. KELLY, D. H. Visual contrast sensitivity. *Op. Acta* 24: 107–129, 1977.
75. KELLY, D. H. Motion and vision. II. Stabilized spatio-temporal threshold surface. *J. Opt. Soc. Am.* 69: 1340–1349, 1979.
76. KELLY, D. H. AND R. E. SAVOIE. A study of sine-wave contrast sensitivity by two psychophysical methods. *Percept. Psychophys.* 14: 313–318, 1973.
77. KELLY, D. H., AND D. VAN NORREN. Two-band model of heterochromatic flicker. *J. Opt. Soc. Am.* 67: 1081–1091, 1977.
77a. KÖNIG, A. Über den menschlichen Sehpurpur und seine Bedeutung für das Sehen. *S. B. Akad. Wiss. Berlin* 559–575, 1894.
78. KRANTZ, D. H. Measurement theory and qualitative laws in psychophysics. In: *Contemporary Developments in Mathematical Psychology. Measurement, Psychophysics, and Neural Information Processing,* edited by D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes. San Francisco: Freeman, 1974, vol. II, p. 160–199.
79. KRANTZ, D. H. Color measurement and color theory. I. Rep-

resentation theorem for Grassman structures. *J. Math. Psychol.* 12: 283–303, 1975.
80. KRANTZ, D. H. Color measurement and color theory. II. Opponent colors theory. *J. Math. Psychol.* 12: 304–327, 1975.
81. KUFFLER, J. W., AND J. G. NICHOLLS. *From Neuron to Brain.* Sunderland, MA: Sinauer Assoc. 1976.
82. LAND, E. H. Color vision and the natural image. Parts I and II. *Proc. Natl. Acad. Sci. USA*, 45: 115–129, 639–644, 1959.
83. LATCH, M., AND P. LENNIE. Rod-cone interaction in light adaptation. *J. Physiol. London* 269: 517–534, 1977.
84. LE GRAND, Y. *Space and Form Vision.* Bloomington: Indiana Univ. Press, 1967.
85. LE GRAND, Y. *Light, Colour, and Vision* (2nd ed.). London: Chapman and Hall, 1968.
86. LENNIE, P., AND D. I. A. MACLEOD. Background configuration and rod threshold. *J. Physiol. London* 233: 143–156, 1973.
87. LIGHTHILL, M. J. *Introduction to Fourier Analysis and Generalized Functions.* Cambridge: Cambridge Univ. Press, 1964.
88. LYTHGOE, R. J. The mechanism of dark adaptation. *Br. J. Opthalmol.* 24: 21–43, 1940.
89. LUDVIGH, E., AND E. F. MCCARTHY. Absorption of visible light by refractive media of the human eye. *Arch. Ophthalmol.* 20: 37–51, 1938.
89a. MACADAM, D. L. *Sources of Color Science.* Cambridge, MA: MIT Press, 1970.
90. MACLEOD, D. I. A. Visual sensitivity. *Annu. Rev. Psychol.* 2: 613–645, 1978.
91. MAFFEI, L., AND F. W. CAMPBELL. Neurophysiological localization of the vertical and horizontal visual coordinates in man. *Science* 167: 386–387, 1970.
92. MAKOUS, W., AND R. BOOTHE. Cones block signals from rods. *Vision Res.* 14: 285–294, 1974.
93. MAKOUS, W., AND D. PEEBLES. Rod-cone interaction: reconciliation with Flamant and Stiles. *Vision Res.* 19: 695–698, 1979.
94. MARC, R. E., AND H. G. SPERLING. Chromatic organization of primate cones. *Science* 196: 454–456, 1977.
95. OSTERBERG, G. Topography of the layer of rods and cones in the human retina. *Acta Opthalmol. Suppl.* 6: 11–97, 1935.
96. PEARSON, D. E. *Transmission and Display of Pictorial Information.* New York: Wiley, 1975.
97. PENN, R., AND W. A. HAGINS. Kinetics of the photocurrent of retinal rods. *Biophys. J.* 12: 1073–1094, 1972.
98. POLYAK, S. L. *The Vertebrate Visual System.* Chicago: Univ. of Chicago Press, 1957.
99. POPPEL, E., R. HELD, AND J. E. DOWLING. Neuronal mechanisms in visual perception. *Neurosci. Res. Program Bull.* 7, 15, 3. Cambridge, MA: MIT Press, 1977.
100. PUGH, E. N., JR. Rhodopsin flash photolysis in man. *J. Physiol. London* 248: 393–412, 1975.
101. PUGH, E. N., JR. Rushton's paradox: rod dark adaptation after flash photolysis. *J. Physiol. London* 248: 413–431, 1975.
101a. RATLIFF, F. *Mach Bands: Quantitative Studies on Neural Networks in the Retinal.* San Francisco, CA: Holden-Day, 1965.
102. RIPPS, H., M. SHAKIB, AND E. D. MACDONALD. Peroxidase uptake by photoreceptor terminals of the skate retina. *J. Cell Biol.* 70: 86–96, 1976.
103. RIPPS, H., AND R. A. WEALE. Contrast and border phenomena. In: *The Eye. Visual Function in Man*, edited by H. Davson. New York: Academic, 1976, vol. 2A, p. 133–184.
104. RIPPS, H. AND R. A. WEALE. The visual photoreceptors. In: *The Eye. Visual Function in Man*, edited by H. Davson. New York: Academic, 1976, vol. 2A, p. 5–41.
105. ROBSON, J. G. Receptive fields: neural representations of the spatial and intensive attributes of the visual image. In: *Handbook of Perception. Seeing*, edited by E. C. Carterette and M. P. Friedman. New York: Academic, 1975, vol. 5, p. 82–116.
106. RODIECK, R. W. *The Vertebrate Retina.* San Francisco, CA: Freeman, 1973.
107. ROSE, A. The sensitivity performance of the human eye on an absolute scale. *J. Opt. Soc. Am.* 38: 196–208, 1948.
108. ROSE, A. *Vision: Human and Electronic.* New York: Plenum, 1973.
109. RUSHTON, W. A. H. The difference spectrum and the photosensitivity of rhodopsin in the living human eye. *J. Physiol. London* 134: 11–29, 1956.
110. RUSHTON, W. A. H. Bleached rhodopsin and visual adaptation. *J. Physiol. London* 181: 645–655, 1965.
111. RUSHTON, W. A. H. The Ferrier lecture. Visual adaptation. *Proc. R. Soc. London Ser. B* 162: 20–46, 1965.
112. RUSHTON, W. A. H. Review lecture: pigments and signals in colour vision. *J. Physiol. London* 220: 1–31, 1972.
113. RUSHTON, W. A. H., AND G. WESTHEIMER. The effect upon the rod threshold of bleaching neighbouring rods. *J. Physiol. London* 16: 318–329, 1962.
114. SAVOIE, R. E. The Bezold-Brucke effect and visual non-linearity. *J. Opt. Soc. Am.* 63: 1253–1261, 1973.
115. SCHADE, O. H. Optical and photoelectric analog of the eye. *J. Opt. Soc. Am.* 46: 721–739, 1956.
116. SCHADE, O. H. *Image Quality. A Comparison of Photographic and Television Systems.* Princeton, NJ: RCA Laboratories, 1975.
117. SNYDER, A. W., AND R. MENZEL (editors). *Photoreceptor Optics.* New York: Springer-Verlag, 1975.
118. SNYDER, A. W., S. B. LAUGHLIN, AND D. G. STAVENGA. Information capacity of eyes. *Vision Res.* 17: 1163–1175, 1977.
119. STABELL, U., AND B. STABELL. The effect of rod activity on colour matching functions. *Vision Res.* 15: 1119–1125, 1975.
120. STILES, W. S. *Mechanisms of Colour Vision.* New York: Academic, 1978.
121. STILES, W. S., AND B. H. CRAWFORD. Equivalent adaptation levels in localized retinal areas. In: *Report of a Joint Discussion on Vision, Physical Society of London.* London: Cambridge Univ. Press, 1932.
122. STILES, W. S., AND B. H. CRAWFORD. The luminous efficiency of rays entering the pupil at different points. *Proc. R. Soc. London Ser. B* 112: 428–450, 1939.
123. TELLER, D. Y., D. P. ANDREWS, AND H. B. BARLOW. Local adaptation in stabilized vision. *Vision Res.* 6: 701–705, 1966.
124. THOMAS, J. P. Spatial resolution and spatial interaction. In: *Handbook of Perception. Seeing*, edited by E. C. Carterette and M. P. Friedman. New York: Academic, 1975, vol. 5, p. 233–264.
126. VAN NESS, F. L., AND M. A. BOUMAN. Spatial modulation transfer in the human eye. *J. Opt. Soc. Am.* 57: 401–406, 1967.
127. WAGNER, G., AND R. M. BOYNTON. Comparison of four methods of heterochromatic photometry. *J. Opt. Soc. Am.* 62: 1508–1515, 1972.
128. WALD, G. Carotenoids and the visual cycle. *J. Gen. Physiol.* 19: 351–372, 1935.
129. WALD, G. Pigments of the bull frog retina. *Nature* 136: 382, 1935.
130. WALD, G. On the mechanism of the visual threshold and visual adaptation. *Science* 119: 887–892, 1954.
131. WALD, G. The receptors of human color vision. *Science* 145: 1007–1017, 1964.
132. WALD, G. Blue-blindness in the normal fovea. *J. Opt. Soc. Am.* 57: 1289–1301, 1967.
133. WALD, G., AND GRIFFIN, D. The change in refractive power of the human eye in dim and bright light. *J. Optical Soc. Am.* 37: 321–336, 1947.
134. WEALE, R. A. Observations on photochemical reactions in living eyes. *Br. J. Opthalmol.* 41: 461–474, 1957.
135. WEALE, R. A. Further studies of photo-chemical reactions in living human eyes. *Vision Res.* 1: 354–378, 1962.
136. WEALE, R. A. Photo-chemical changes in the dark-adapting human retina. *Vision Res.* 2: 25–33, 1962.
137. WERBLIN, F. S. Adaptation in a vertebrate retina: intracellular recording in *Necturus*. *J. Neurophysiol.* 34: 228–241, 1971.
138. WESTHEIMER, G. Modulation thresholds for sinusoidal light distributions on the retina. *J. Physiol. London* 152: 67–74,

1960.
139. WESTHEIMER, G. Bleached rhodopsin and retinal interaction. *J. Physiol. London* 195: 97–106, 1968.
140. WESTHEIMER, G. Visual acuity and spatial modulation thresholds. In: *Handbook of Sensory Physiology. Visual Psychophysics,* edited by D. Jameson and L. M. Hurvich. Berlin, Springer-Verlag, 1972 vol. 7, pt. 4, p. 170–187.
141. WILLIAMS, D. R., D. I. A. MACLEOD, AND M. M. HAYHOE. Punctate sensitivity of the blue-sensitive mechanism. *Vision Res.* 21: 1357–1375, 1981.
142. WILSON, H. R., AND J. R. BERGEN. A four mechanism model for threshold spatial vision. *Vision Res.* 1: 19–32, 1979.
143. WINSOR, C. P., AND A. CLARK. Dark adaptation after varying degrees of light adaptation. *Proc. Natl. Acad. Sci. USA* 22: 400–404, 1936.
144. WYSZECKI, G., AND W. S. STILES. *Color Science.* New York: Wiley, 1967.
145. YELLOTT, J. I., JR. Spectral analysis of spatial sampling by photoreceptors: topological disorder prevents aliasing. *Vision Res.* 22: 1211–1218, 1982.
146. YOUNG, T. Note on a paper by Dalton. In: *Lectures on Natural Philosophy,* (1st ed.), 1807, vol. 2, p. 315.