MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. MEMO No. 764 1 May 1984

COLOR VISION: REPRESENTING MATERIAL CATEGORIES

John M. Rubin and W.A. Richards

Abstract. We argue that one of the early goals of color vision is to distinguish one kind of material from another. Accordingly, we show that when a pair of image regions is such that one region has greater intensity at one wavelength than at another wavelength, and the second region has the opposite property, then the two regions are likely to have arisen from distinct materials in the scene. We call this material change circumstance the "opposite slope sign condition." With this criterion as a foundation, we construct a representation of spectral information that facilitates the recognition of material changes.

Our theory has implications for both psychology and neurophysiology. In particular, Hering's notion of opponent colors and psychologically unique primaries, and Land's results in two-color projection can be interpreted as different aspects of the visual system's goal of categorizing materials. Also, the theory provides two basic interpretations of the function of double-opponent color cells described by neurophysiologists.

Acknowledgment. This report describes research done at the Department of Psychology and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology, and was supported by NSF and AFOSR under a combined grant for studies in Natural Computation, grant 79-23110-MCS, and by the AFOSR under an Image Understanding contract F49620-83-C-0135. John Rubin was supported by an NSF Graduate Fellowship, and by a pre-doctoral fellowship from the M.I.T. Center for Cognitive Science. The authors would like to thank T. Poggio, Nancy Kanwisher, Andrew Knapp, and the members of the Natural Computation group for their comments, and Bill Gilson for a meticulous reading of the manuscript.

© 1984 Massachusetts Institute of Technology

1. Introduction

The human visual system performs a remarkable feat. The pattern of light that reaches the eye from a scene is the result of a complex interaction among several factors: the quality of the illuminant, the geometry of the scene, and the properties of the materials composing the visible surfaces. Yet somehow these confounded factors are mostly separated in our perception. We see particular spatial arrangements of objects. These objects appear bounded by surfaces having properties—color and texture—roughly invariant over a range of conditions of geometry and illumination. To compute invariant descriptions of the material properties of surfaces is an important goal of any visual system. Such material descriptors are useful for object recognition and visual search.

It's commonplace to assume color vision has something to do with capturing the albedoes of surface materials.[1] But exactly what aspect of the albedo function would serve a visual system best? Consider the grandiose goal of recovering a material's albedo as a continuous function of wavelength. Not only is this goal impractical; it is counter to the aim of finding invariant descriptors. With such an over-zealous representation, unimportant variations in a surface would prevent its being recognized as a single region, a patch of one kind of stuff. The perception of the world would be shattered with spectral acuity too fine; one literally wouldn't be able to see the forest for the trees.

Here we seek a representation of material reflectance in which trivial surface variations can be overlooked in order to appreciate important similarities.[2] At the same time, the representation must allow some discrimination among different materials. Below we develop such a categorical color space, based on a theoretical solution to the problem of identifying material changes. A trichromatic system, it will be shown, yields a two-dimensional color space in which the axes will turn out to represent boundaries between different materials. The four quadrants of the two-dimensional space represent material categories.

2. Spectral Information at Edges

When two image regions arise from different materials in the scene, the transition from one material to another will usually bring about an edge in the image. Thus we restrict our search for material changes to edges. How can we decide whether an edge is due to a material change?

An edge in the image will usually arise from a single event or state of affairs in the three-dimensional scene (Marr, 1982). The most common edge types are shadows,

[1] The albedo of a material is a function of wavelength $\rho(\lambda)$, with range $(0, 1)$, that indicates what fraction of photons (emitted by some light source) at each wavelength will be reflected.

[2] We are not suggesting any spectral information be thrown away. We are merely exploring a single problem. Other problems may require detailed spectral information.

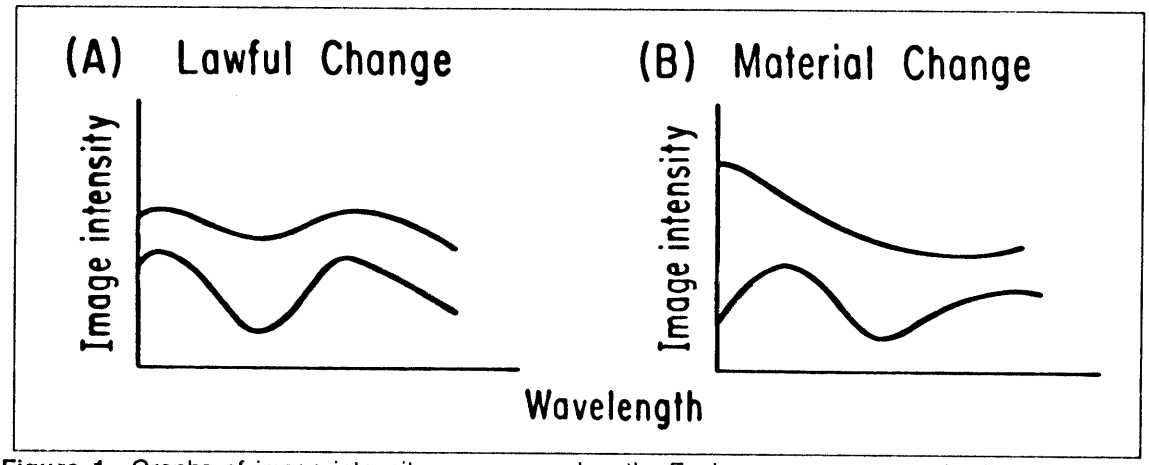

Figure 1 Graphs of image intensity versus wavelength. Each curve represents the image intensity measurable from one image region. A) Two graphs of same shape: a likely lawful change. B) Two graphs of different shape: a candidate for material change.

highlights, surface orientation discontinuities, and pigment density changes.[3] Alternatively, an edge may be due to a material change, a discontinuity between two different kinds of stuff.[4] How can a material change edge be distinguished from other types of edges? Rubin & Richards (1982) attempted to answer this question. Edges which arise from shadows, orientation changes and highlights are *lawful* in the sense that there are equations that describe how image intensities will change across these edges. By contrast, material changes are completely unpredictable; they are arbitrary changes, and as such, can only be inferred by ruling out, at a given edge, the possibility of any of the above lawful changes.

To infer material changes, we now face the awkward prospect of having to reject, one by one, each of the lawful changes. Perhaps there is some method of rejecting all of those edges *en masse*. Fortunately, there is a simple ordinal rule common to all the edges formed by lawful processes: if the intensity at one wavelength decreases across a lawful edge (shadows, highlights, and so on) then the intensity must also decrease at all other wavelengths taken across the same edge (Rubin and Richards, 1982). When this condition is violated, we say there is a "spectral crosspoint" across the edge. Spectral crosspoints imply material changes; a spectral crosspoint is illustrated in Fig. 2a. The spectral crosspoint is not the only means of discovering material changes, however. We will show that a second and independent condition holds for each of the lawful processes—namely the preservation of ordinality of image intensity across wavelength. A violation of this condition implies a material change.

[3]Surface orientation change and shadow can coincide at an edge, but this exception is unimportant to the arguments that follow. See Rubin & Richards, 1982, footnote 16.

[4]We consider materials to consist of some spectrally neutral embedding material (e.g., cellulose) impregnated with a single pigment (e.g., chlorophyll). A material change is a change in pigment type, or a change in both pigment and embedding material.

3. The Opposite Slope Sign Inference

3.1 The Lawful Processes

Figure 1a shows two image intensity graphs of the same shape. Intuitively, the two graphs, of similar shape, arise from measurements taken on either side of a "lawful" edge type. Figure 1b shows two graphs of different shape. None of the lawful edge types could have produced such a distortion, and intuitively it seems that a material change edge is the best explanation. We now must make explicit what we mean by "same shape" and then show that this definition of spectral shape remains invariant across edges created by shadows, changes in surface orientation, highlights or variations in pigment density—namely the lawful conditions we wish to reject as material changes.

> *Definition:* Two curves of intensity versus wavelength have the same shape if the ordinal relations of image intensity across wavelength are preserved.

Thus, if $I_X(\lambda)$ and $I_Y(\lambda)$ are image intensities as functions of wavelength measured on both sides, X and Y, of an edge. $I_X(\lambda)$ and $I_Y(\lambda)$ have identical ordinality if, for all λ_1 and λ_2, $I_X(\lambda_1) < I_X(\lambda_2)$ iff $I_Y(\lambda_1) < I_Y(\lambda_2)$. Note that two image intensity functions of identical ordinality will have local extrema at the same values of wavelength.

Given this ordinal definition of "same shape", Appendix 1 shows that the ordinality relationship is preserved across all edges arising from the lawful edge types, provided that the following two conditions hold:

> *Gray world condition:* The average of all the different albedoes in the scene will be a spectrally flat "gray", so that the diffuse reflected light will have the same spectral character as the direct light.

> *Spectral normalization:* The spectral samples of image intensity have been normalized with respect to the color of the illuminant.

(The need for the second condition, namely spectral normalization, will be eliminated subsequently.)

3.2 The Opposite Slope Sign Operator

We now can proceed to test for "same shape" using the ordinality relation. If ordinality is violated across an edge, then we infer the edge does not arise from one of the "lawful"

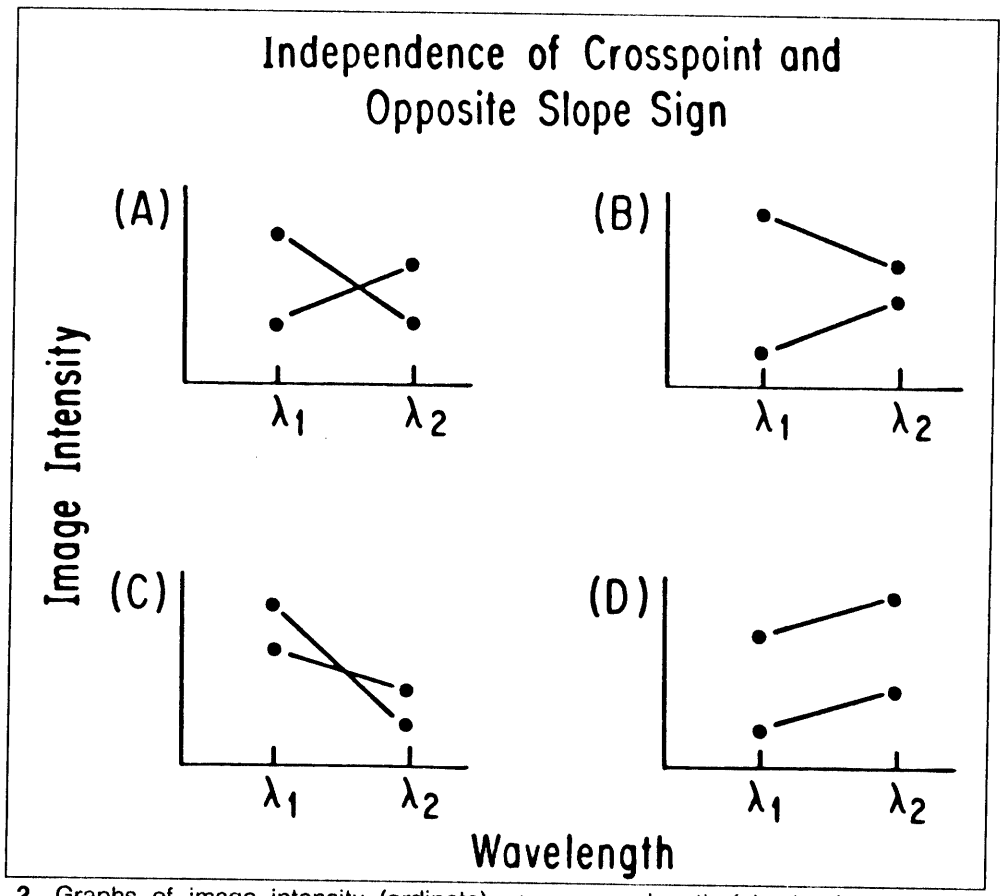

Figure 2 Graphs of image intensity (ordinate) versus wavelength (abscissa). Two wavelength samples, λ_1 and λ_2, are shown. An image region yields two samples of intensity, one for each wavelength, and is represented by the line segment connecting the two sample values. a) & c) Two examples of the spectral crosspoint (Rubin & Richards, 1982). a) & b) Two examples of the opposite slope sign condition. This is the minimal configuration that shows different ordinalities. Note that the crosspoint and opposite slope sign condition are completely independent, since they can occur together (a), or each can occur alone (b and c), or neither can occur (d).

processes and hence must represent a material change (provided also, of course, that our grey world condition is not violated).[5]

What is the simplest way to seek violations of ordinality? A pair of spectral samples suffices. Let the image intensities on both sides of an edge be measured at wavelengths λ_1 and λ_2. If image intensity at λ_1 is greater than that at λ_2 on one side of the edge, then the ordinality condition requires the same relationship hold on the other side. So if the two sides of the edge do not have greater intensity in the same spectral sample, ordinality is violated; the edge cannot be lawful. (Details are given in Appendix 1.) This condition

[5]It is possible when the grey world assumption is wrong, material changes will be inferred from images. This is not entirely bad news; if human perception also goes awry when the grey world assumption is violated, then our theory will become more credible as an account of biological visual systems.

is called the *opposite slope sign condition.*[6] Examples are shown in Fig. 2a and 2b. The "slope" of the opposite slope sign condition is the slope of the graph of intensity versus wavelength; it is an evaluation of the sign of the derivative of the spectral image intensity function, $\frac{dI}{d\lambda}$.

More formally, given two regions X and Y across an edge and intensity samples I taken at two wavelengths λ_1 and λ_2, we have the following test for a material change:

Opposite Slope Sign Condition:

$$(I_{X\lambda_2} - I_{X\lambda_1})(I_{Y\lambda_2} - I_{Y\lambda_1}) < 0.$$

which may be contrasted with the previously derived crosspoint condition (Rubin and Richards, 1982):

Spectral Crosspoint Condition:

$$(I_{X\lambda_1} - I_{Y\lambda_1})(I_{X\lambda_2} - I_{Y\lambda_2}) < 0.$$

Note that the spectral crosspoint and the opposite slope sign conditions are completely independent. Figure 2a shows the two occurring together. Each condition can arise alone, as shown in Figs. 2b and 2c. Finally neither condition is necessary, as shown in Fig. 2d.

The two conditions are related by a kind of symmetry. The spectral crosspoint must make two comparisons across an edge (one for each wavelength), and combine them logically (both comparisons must work out in the correct way). The opposite slope sign condition must make two comparisons, one within each image region, and then combine them logically across the edge.

To summarize: the spectral crosspoint—our original means of finding material changes—has been augmented by a second and independent material change condition: opposite slope sign. The opposite slope sign condition is the key theoretical result on which we will base our spectral representation of material types. We choose opposite slope sign rather than the crosspoint, because the opposite slope sign condition tells us something about each of the two regions that produce it. Namely, one region has positive spectral slope, the other negative. By contrast, the spectral crosspoint cannot be decomposed into assertions about the two regions that produce it. In a crosspoint, spatial and spectral information are

[6]The opposite slope sign condition is described here as existing statically, across an edge. It is a spatial comparison of spectral information. A comparison of spectral information in time is equivalent. Such a *temporal opposite slope sign condition* would work as follows: An eye could sweep across an edge, and the spectral information before and after the movement could be compared. Similarly, there is a temporal equivalent of the crosspoint. Consequences of these isomorphic computations in the temporal domain will not be explored here.

hopelessly intertwined. We do not cast aside the crosspoint, though. It will play a vital role in correcting for the spectral content of the illuminant.

4.0 Spectral Normalization

For the opposite slope sign test to find material edges successfully, it is necessary for the measured spectral intensities to be normalized. That is, these samples must be transformed to what they would have been under a spectrally flat ("white") illuminant. Clearly if no correction is applied, then the stronger spectral skew of an illuminant may not only reduce the number of observed opposite slope sign pairs, but more seriously, may transform pairs having the same slope sign into pairs that are seen as having an opposite slope sign.

By contrast, the spectral crosspoint condition is insensitive to the spectral content of the illuminant, as can be seen by inspecting panels A and C of Fig. 2. (See Rubin & Richards, 1982, for a more formal treatment.) We capitalize on this property of the crosspoint to devise a theory of spectral normalization. Once the image has been spectrally normalized, it is as if the illuminant were white. The opposite slope sign condition will now be able to find correctly a maximum number of material changes.

Consider now a scene composed of a large number of randomly selected materials. For each image region (simple closed curves defined by edges), take two samples of intensity I_{λ_1} and I_{λ_2} at wavelengths λ_1 and λ_2. Each region will be associated with a spectral slope sign, which is just the sign of the difference $I_{\lambda_1} - I_{\lambda_2}$. If the illuminant were white (same photon flux at all wavelengths), we would expect to have roughly equal numbers of regions of positive spectral slope and regions of negative spectral slope. This expectation is based on two assumptions. The first is that there is a random collection of materials in the scene. The second is that materials in the world are such that a random collection of them will be divided equally between positive and negative spectral slope.

As suggested above, normalization requires a collection of image regions that arises from a random set of materials. What about using *all* image regions? The set of all image regions is not likely to represent a random collection of materials, because many materials will recur in several image regions. For example, if a cast shadow cuts across a single piece of material, that material will be twice represented, once for each side of the shadow edge. A second example arises with pigment density changes. In a forest scene, all leaves are composed of the same material (chlorophyll embedded in a cellulose base). A sensible normalization scheme would not take each leaf as a distinct patch of material; minor variations in pigment density from leaf to leaf ought to be ignored.

It seems clear, then, that not all image regions should participate in normalization. Perhaps a subset of image regions can be found that is more likely to represent a random collection of materials. The spectral crosspoint offers a means of finding such a random subset of regions. Suppose that instead of taking each image region as a distinct material, we took only pairs of regions that have a spectral crosspoint on the edge between them. We would be guaranteed that each pair of regions would correspond to distinct materials. The pairs of different material regions found with the crosspoint will be the subset of image regions that will be used for normalization.

Our normalization scheme works like this: Recall that we expect the regions found by the crosspoint to represent a random collection of materials. So we expect roughly the same number of regions having positive spectral slope as negative. For the subset of image regions defined by the crosspoint, tally the number having positive spectral slope and the number having negative slope. If the numbers are approximately equal, our expectation has been met; we can infer that the illuminant is white (spectrally flat).[7] Suppose to the contrary that the number of regions of positive spectral slope exceeds the number of negative-slope regions. Then we can infer that the illuminant is more intense at long wavelengths than at short. (Positive spectral slope means greater intensity in the longer wavelength sample.) Now multiplicatively scale one of the spectral samples. In the example here, we need to multiply all long wavelength samples by some number less than one. Exactly which number? The one that will fulfill our expectation of equal numbers of positive and negative spectral slope. That is, multiply all long wavelength samples by some number (less than one) such that half of the regions under consideration will have greater intensity in the modified long wavelength sample than the short wavelength sample, and half, the reverse. For a large number of samples, the multiplicative constant of normalization can be calculated from the mean value of the spectral slopes of all regions participating in crosspoints. See the algorithm for spectral normalization in Appendix 2.

This crosspoint normalization scheme has some useful properties. Each image region used has the same potency in normalization, regardless of the size of the region. That is, each pair of image regions (found with the crosspoint) maps to a pair of data points, one for each region. This is good for two reasons. First, the scheme is independent of image region areas. This is desirable since we would not want visual systems to treat an image of a large blue thing and a small red thing differently from an image of a small blue thing and

[7]Note there must be some crosspoints for normalization to proceed. If there are no crosspoints, there are no regions to consider. So although it is technically true that there are equal numbers of positive-slope regions and negative-slope regions (namely, zero), we do not want to infer the illuminant is white for two reasons. First, we have no information about any image region, and thus it seems imprudent to guess blindly that the light is white. Second, we have evidence that the scene consists of a single material since it has no crosspoints. Normalization would bring about material change assertions via the opposite slope sign condition, in contradiction to the evidence of uniformity from the crosspoint.

It is worth comparing our crosspoint normalization with Land's latest normalization theory. Land's (1983) scheme involves comparing the image intensity of a target region with that of a few hundred *random* locations in the image. In such a theory, the larger an image region, the more random locations it will contain. Land's theory is therefore area-based, while ours is independent of the particular sizes of image regions. Our theory makes different predictions from Land's: we expect no effect on normalization from the sizes of image regions, or from the lengths of image edge segments.

5.0 Choosing a Representation

Assume now that the image has been normalized using the spectral crosspoint condition, as described in section 4. We next select a representation of spectral information based on that rule. In particular, *we seek a simple, convenient spectral representation of materials that is invariant under shadow, highlight, surface orientation change, and pigment density change.*

For any region in the image, intensity can be measured at a long wavelength and at a second, shorter wavelength. Call these two measurements of image intensity L and S, respectively, for each image region. Suppose we'd like to represent the spectral character of a region with a single number, namely some mapping of the pair (L, S). Furthermore, we would like the mapping (L, S) to be invariant under the lawful changes. The recognition of material differences would be easy in such a representation. A single material in its different guises—fully lit, shadowed, having different densities of pigmentation, with different surface orientations—would map ideally to a single point. If there were such a mapping, then whenever two image regions mapped to distinct points, we would know they corresponded to distinct materials.

The lawful edge types are unfortunately so diverse that there is no function giving us the desired mapping. No single continuous function of (L, S) will be invariant under multiplicative (shadow), exponential (pigment density), and additive (highlight) changes. Material change, then, cannot be reduced to the problem of distinguishing two points in the range of some function.

The problem isn't hopeless, however, for there is a continuous function invariant under *some* of the lawful changes, namely the multiplicative ones (shadow and surface orientation change). Consider again the two image intensity samples S and L. The quotient $\frac{L}{S}$ will have the identical value on both sides of a surface orientation change or a shadow edge. The simple quotient is, of course, not unique in remaining constant across an orientation edge. Many functions of the two samples L and S have the same property. We will choose among three simple functions having this property:

Many functions of the two samples L and S have the same property. We will choose among three simple functions having this property:

$$\frac{L}{S} \qquad \frac{L}{L+S} \qquad \frac{L-S}{L+S} \tag{7}$$

How can we select among these candidates? The function $\frac{L}{S}$ takes image regions into the unbounded interval $(0, \infty)$, while the other two functions take image intensities into closed intervals. ($\frac{L}{L+S}$ maps intensities into $[0, 1]$; $\frac{L-S}{L+S}$ maps into $[-1, 1]$.) The function $\frac{L}{S}$ will be rejected, since any reasonable computational system will be better off using quantities that fall within a closed interval, rather than those that could be arbitrarily large. To choose between the two remaining candidate functions we consider the ease of discovering material changes in these two maps. In particular, how does the opposite slope sign condition appear in each of the candidate mappings?

Given two image regions X and Y, let F denote the function $\frac{L}{L+S}$, so that $F(X)$ and $F(Y)$ are the values of the function F of regions X and Y, respectively. Then for F, the opposite slope sign condition is expressed by $[sign(F(X)-\frac{1}{2}) \neq sign(F(Y)-\frac{1}{2})]$. (The reason for this expression is that the function F takes on the value $\frac{1}{2}$ whenever $L = S$.)

Let G denote the function $\frac{L-S}{L+S}$, a common measure of contrast. This is a simple function that facilitates the computation of material change. The sign of G is the sign of the spectral slope of an image region. That is, $[sign(G(X)) \neq sign(G(Y))]$ emerges as the opposite slope (material change) condition.

We prefer the function G to the F for our representation. Whereas to determine material change with G requires only a sign check, with F, the system must maintain the constant $\frac{1}{2}$ and perform two subtractions. The particular choice of F or G, though, seems not to be critical for the goals we have in mind.

Figure 3 shows the interval $[-1, 1]$, the range of the function G. Two image regions corresponding to lit and shadowed versions of the same material, or two different surface orientations, will, by design of G, be mapped to the same point. This is shown in Fig. 3a. Two image regions of different pigment density have the same slope sign; hence, in the G map, the corresponding pair of points cannot straddle the zero. The same holds for a pair of points corresponding to a highlight and a neighboring matte region. The latter two edge types are shown in the G mapping in Fig. 3b. If two image regions are mapped to points straddling the zero (Fig. 3c), they arise from different materials.

To summarize, we sought a function of spectral information invariant over the lawful changes. That goal being impossible, we chose $\frac{L-S}{L+S}$ for two reasons. First, it is invariant across shadows and surface orientation changes. Second, finding material changes with the opposite slope sign condition is easy. The range of the function can be divided into

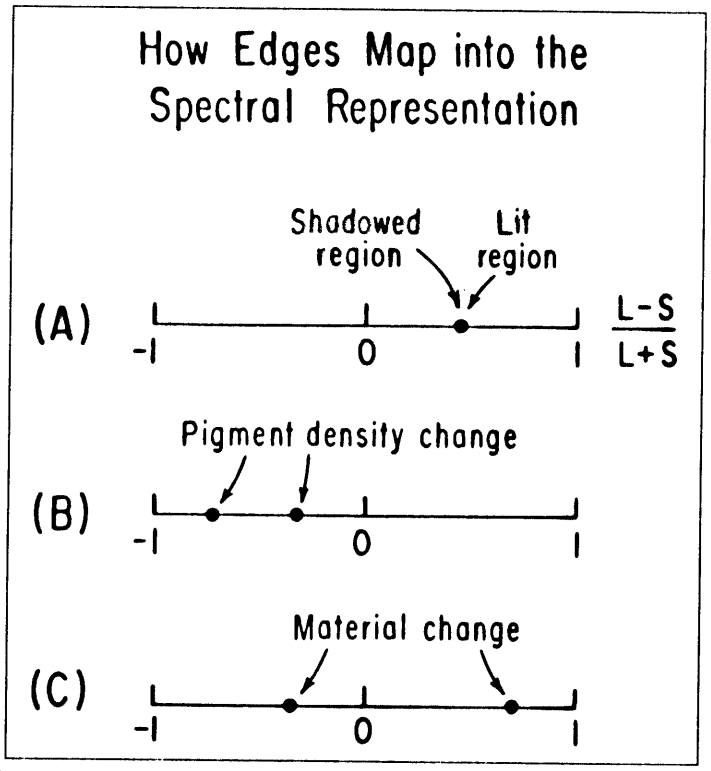

Figure 3 How various processes appear in the spectral representation implied by the mapping $\frac{L-S}{L+S}$, the range of which is [-1,1]. a) Two image regions differing only in surface orientation or shadow map to a single point. b) Two regions differing as matte and highlighted, or as two different degrees of pigmentation density, map to the same half of the range, i.e., they map to points having same-sign coordinates. c) Only two different materials can map to points straddling the zero, i.e., to points of different-sign coordinates.

two parts, $(-1, 0)$ and $(0, 1)$. Materials with albedoes of positive spectral slope sign will map into the positive half of the range, and negative-sloping albedoes to the negative part of the range.[8]

Finally, it's worth reiterating why we built our spectral representation around the opposite slope sign condition, and not the spectral crosspoint. Spectral slope sign is an invariant property of a material's albedo function.[9] The opposite slope sign condition can be decomposed into separate meaningful statements about properties of two image regions: The slope sign of one region is positive, and that of the other, negative. We know something about each region. The crosspoint, by contrast, hopelessly confounds spatial and spectral information. Higher goals of color vision involve describing the properties of individual image regions, and cannot be reached by the crosspoint alone.

[8]Many continuous maps share the same invariance. We selected our map on the basis of *algorithmic* considerations. The particular choice is independent of the *theory* of finding material change edges.

[9]Since a material is defined as a kind of stuff, a single material can have different albedoes as pigment density changes. What stays constant over these changes in density of pigment is spectral slope sign.

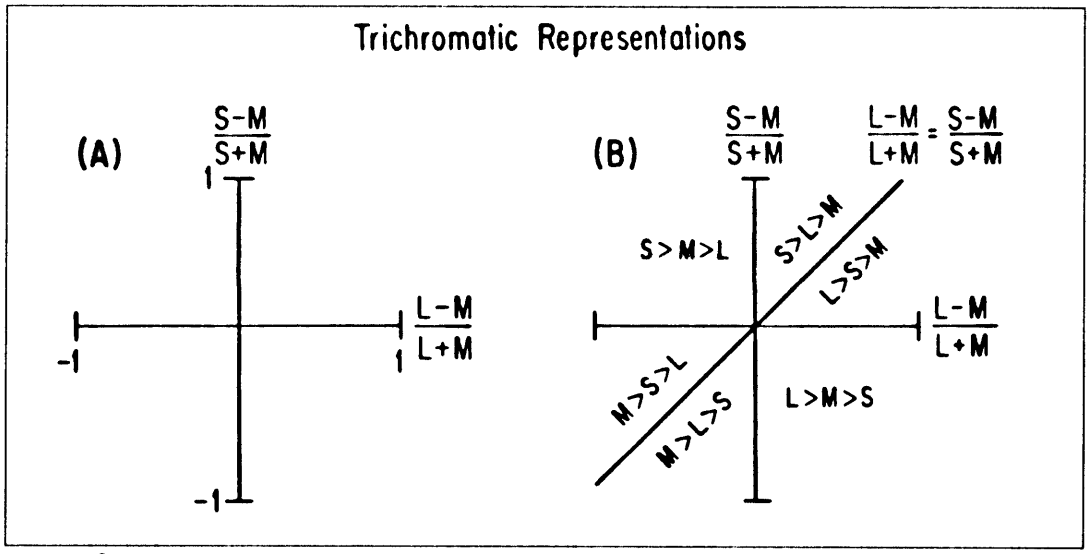

Figure 4 Steps in the construction of the trichromatic material representation. a) Two axes comparing L and M, and S and M samples, are joined orthogonally. Each quadrant is a material category. Points in different quadrants correspond to distinct materials. Points within one quadrant may belong to the same material; they are considered equivalent in this representation. b) The line of unit slope in the figure above represents the comparison between S and L samples. Adding the unit slope line divides the color space into six regions or "hextants." Points in different hextants arise from different materials. Note the hextants do not have equal areas.

6.0 Trichromacy: Finding More Material Changes

Suppose we add a third spectral sample, call it M, to our original S and L samples. Adding a third spectral sample will allow the detection of new kinds of material changes.[10] However, more importantly, the number of basic material categories will be increased from two to six.

In the two-wavelength-sample material representation, an image region is encoded essentially by the rank order of the spectral samples, or equivalently by the sign of the slope of the line segment connecting the samples. Thus, given two wavelength measurements, there are two types of material—negative slope and positive slope. With three wavelength samples, an image region is associated with three slope signs—a slope between each pair of samples (SM, ML, SL). There are six possible rank orderings of the measurements ($3! = 6$), and thus six possible basic material types. Any two regions that produce distinct rank orderings of the wavelength samples will bring about one or more opposite slope signs. Any two such regions must therefore be distinct materials.

As a first step in constructing the trichromatic material representation, we combine slope information from two of the three pairs of samples. Arbitrarily, we begin with SM

[10]The additional number of material changes detected with each new spectral sample will drop sharply after the third sample. The reason is that the albedoes of natural objects (in the visible range) are typically slow-changing functions of wavelength (Krinov, 1971; Snodderly, 1979). Cohen (1964) showed that three carefully chosen functions of wavelength captured over 99% of the albedo functions of Munsell chips.

and LM, combining the information in a two-dimensional space as shown in Fig. 4a. Image regions are mapped to points in the square $[-1,1]$ **X** $[-1,1]$, and a pair of points separated by an axis (or both axes) correspond to two regions of different material, just as did a pair of points straddling the zero in Fig. 3c. Any pair of points in a single quadrant may arise from a single material. This is the sense in which quadrants represent material categories. Without yet considering comparisons between S and L samples, we already have a categorical representation in Fig. 4a, in which in each quadrant corresponds to a material category.

Let's now examine the third pairing of samples, S and L. What condition holding between a pair of points in the preliminary representation of Fig. 4a corresponds to the opposite slope sign condition between S and L? It is easily shown that if a pair of points straddles the line of unit slope, the points arise from materials with opposite (S and L) slopes.[11] Furthermore, not just the sign, but the continuous value $\frac{L-S}{L+S}$ of the L to S comparison is contained implicitly in the representation defined by ordered pairs $(\frac{S-M}{S+M}, \frac{L-M}{L+M})$ that Fig. 4a illustrates.[12]

The unit slope line in the SM–LM space therefore has special significance, and is added to the representation as a third material change axis in Fig. 4b. A pair of points lying across any of the three axes will correspond to distinct materials. Thus, each of the six sectors of Fig. 4b corresponds to a material type, or equivalently, to a rank ordering of the three samples. The particular rank ordering associated with each "hextant" is shown in Fig. 4b. Note the hextants of Fig. 4b do not have equal areas. The original pair of axes can be joined in a skew fashion to allocate more or less area to the different material categories.

To summarize, image intensities are measured at S, M and L, normalized according to the crosspoint normalization of section 4, and mapped to $(\frac{L-M}{L+M}, \frac{S-M}{S+M})$ in a rectangular coordinate system, initially creating four basic material types. A further subdivision into six types can arise by using the line of unit slope as a third axis, dividing the region $[-1,1]^2$ into six regions, each corresponding to a different material type. Points in different hextants arise from different materials, whereas points common to one hextant may arise from lawful edge events occurring on a single material.

Algorithm aficionados should turn to Appendix 2, where we sketch a procedure for spectral categorization based on the above theory.

---

[11] The line of unit slope is given by $\frac{S-M}{S+M} = \frac{L-M}{L+M}$. This is equivalent to $(S-M)(L+M) = (S+M)(L-M)$, or $S = L$. Points above this unit slope line correspond to $L > S$, points below to $S > L$.

[12] Given the values $(\frac{S-M}{S+M}, \frac{L-M}{L+M})$, we can compute the value of $\frac{L-S}{L+S}$. Let $Q = \frac{S-M}{S+M}$ and $R = \frac{L-M}{L+M}$. Then $\frac{L-S}{L+S} = \frac{Q-R}{QR-1}$.

7.0 Relation to Psychophysics and Neurophysiology

Our spectral representation of material types is but an abstract model of biological color vision. In our theory, certain terms are left undefined. We haven't described what the "spectral samples" of the theory are, and we haven't said anything about how materials are encoded. How then can we assess its relevance? Two linking assumptions will guide the interpretation of our theory. First, in the discussion of the psychology of color vision, we will argue that of the traditional color variables hue, saturation, and lightness, it is hue that encodes material type. Second, in the discussion of neurophysiology, we take the small step to identify the spectral samples of our theory with the relative stimulation of the three human cone photopigments (or combinations thereof).[13] Given this interpretation of our theory, it turns out that double-opponent units found in color neurophysiology can be understood as performing the spectral crosspoint and/or the opposite slope sign computation.

7.1 Psychologically Unique Primaries

Ewald Hering (1964) offered a psychological account of human color perception that was based on the notion of opponent processes. He observed that "redness and greenness, or yellowness and blueness are never simultaneously evident in any color, but rather appear to be mutually exclusive." This is a clear case of categorical perception. *Reddish* and *greenish* are mutually exclusive hue categories, and if hue is encoding material properties, then the two categories will partition materials. See Fig. 5a. Similarly, *bluish* and *yellowish* will partition materials. See Fig. 5b. These two sets of mutually exclusive hue pairs divide the color space into four regions, as in Fig. 5c, just as did our trichromatic color space (Fig. 4a).

Our claim that Hering's color quadrants correspond to our material categories is predictive: we expect that shadows, surface orientation changes, and pigment density changes would only rarely cause perceived hue to change from *reddish* to *greenish* (or vice versa), or from *yellowish* to *bluish* (or vice versa). As noted in Appendix I, highlights could be troublesome.

The fact that there are four hue categories supports the idea that trichromatic human vision uses two opposite slope sign checks, as in Fig. 4a, but not the third, as shown in Fig. 4b. (Goethe [1808], however, proposed a theory of color perception based on six hue categories, which might correspond to the use of all three opposite slope sign checks.)

[13]Our theory of crosspoints and opposite slope signs was based on spectral samples *at a single wavelength*. Biological measurements of the spectrum are broadband. It turns out that broadband samples cannot introduce crosspoints that are false targets. That is, a spectral crosspoint found with broadband samples is still a reliable indicator of material change (Rubin & Richards, 1982, Appendix IV). The opposite slope sign condition may not be as robust; more work is needed to study the effects of broadband sampling.

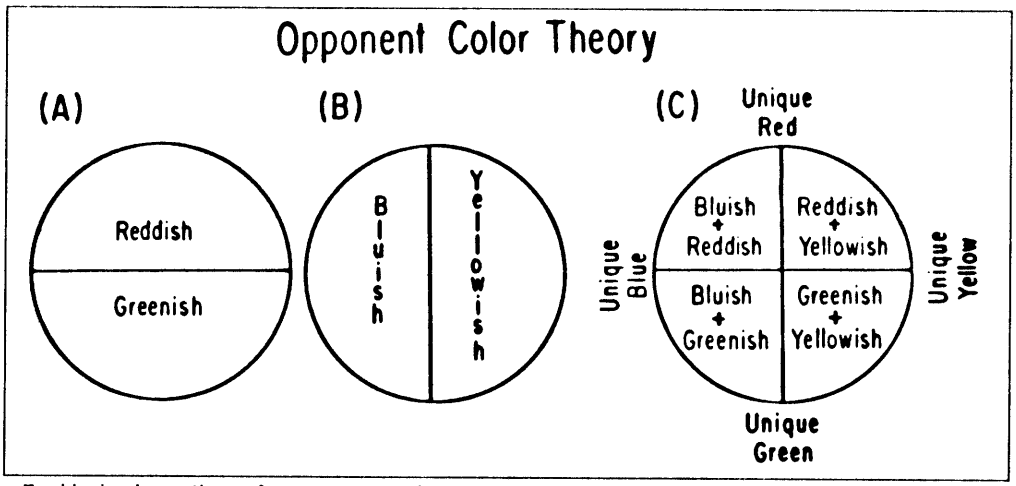

Figure 5 Hering's notion of opponent color processes. a) All colors are either reddish or greenish, but never both. b) All colors are either bluish or yellowish, but never both. c) The two pairs of mutually exclusive colors divide the color circle into four quadrants, similar to the trichromatic representation that we develop in Fig. 6a.

Evidence from infants (Bornstein *et al.*, 1976) supports Hering's theory of four hue categories as independent of language and culture. Pigeons also have categorical color perception (Wright & Cumming, 1971), suggesting the computational scheme that we propose here is fundamental to color vision across species.

Hering's notion of opponent color processes implies four special hues. They are indicated in Fig. 5c. These hues, which Hering called psychological primaries, are the boundaries that separate color categories. Primary red is that hue among the reddish hues that separates the yellowish from the bluish; primary blue is that hue among the bluish that splits the reddish from the greenish; and so on. These primary colors are unstable in the sense that any deviation from them involves a change of color categories. Hering's psychological primaries correspond to the axes of our trichromatic representation (Fig. 4a).

Just why these primaries have their particular locations in the spectrum is an interesting evolutionary question not addressed here. One possibility is that a creature's material boundaries are positioned in some way as to make the greatest number of discriminations among materials encountered in its environment.[14] Interesting work has been done along these lines. Snodderly (1979) attempted to relate the color vision of New World monkeys to the spectral characteristics of their jungle habitat. Levine & MacNichol (1982) and McFarland

[14]Material boundaries can be changed in two ways. The wavelength at which a photopigment captures the greatest percentage of photons can be altered, or new "channels" can be created by combining photopigments. One sort of combination of two spectral samples S and L is a rotation; that is, new coordinates $(S\cos\theta - L\sin\theta, S\sin\theta + L\cos\theta)$ can be created for some angle of rotation θ. The original and rotated coordinate systems will not always agree about whether two image regions satisfy the opposite slope sign condition. That is, the two spectral coordinate systems differing only by a rotation will make different material distinctions. An angle θ can therefore be selected to maximize the number of material changes detected.

& Munz (1975b) linked the photopigment characteristics of fishes to the spectral character of light in their environments.

In sum, our spectral representation of material categories is a two-dimensional space in which each quadrant represents a material type, and the axes represent the boundaries between categories. Image regions that map to different quadrants necessarily arise from distinct materials; image regions that map to the same quadrant may arise from a single material. Supposing that hue encodes material information, Hering's observation about human color vision makes sense: hues are divided into four fundamental categories by the mutually exclusive pairs red–green and blue–yellow.

7.2 Land's Experiments

7.2.1 Two-Color Projection

Edwin Land (1959a,b) conducted some remarkable experiments in two-color projection of natural images. Some of the phenomena he reported can be understood in terms of our "materialistic" theory of categorical color vision.

Land's paradigm was as follows. Two different black-and-white transparencies were made of a colorful natural scene by means of long- and short-wavelength filters.[15] The two transparencies were called the long and short records, respectively. Corresponding regions of the two records, in general, were of different grey values. The two records were projected on a screen in register, the short record with short wavelength light, the long record with long wavelength light. Surprisingly, the resulting image was richly colored and faithful to the original still-life.

Land's (1959a,b) work was basically descriptive. He found a means of predicting the hue name of a region in the two-color reconstruction. The intensity of long-wavelength light in the region was expressed as a fraction of the maximum long-wavelength intensity in the entire image. The same was done for short-wavelength intensity, yielding a pair of numbers (each between 0 and 1). This pair of numbers (fraction of maximum S, fraction of maximum L), plotted on log–log axes, yielded a coordinate system that Land used to relate image intensity to perceived hue. Land's coordinate system (hereafter called "Landspace") is shown in Fig. 6a.

We will now try to relate our current work to Land's findings. Whereas Land began with some surprising *experimental* observations of color appearance, we took image intensity

[15]The transparencies did not consist solely of black regions and white regions, but rather the full range of grey values between black and white.

equations as the starting point of our *theoretical* investigation of the problem of discriminating materials. We will show how these two approaches dovetail.

Our argument below consists of four major points. First, we look at how the spectral crosspoint appears in Landspace. Second, we propose that an absence of crosspoints should cause a total failure of Landspace, and note the failure conditions already observed for Landspace correspond to such an absence. Accordingly, we make some predictions for two-color projection that conflict with predictions in the literature. Third, we note the opposite slope sign condition is identical to the fundamental split between warm and cool color categories in Landspace. Finally, we suggest a straightforward extension of our crosspoint normalization theory that would account for a peculiar result in two-color projection.

7.2.2 The Spectral Crosspoint in Landspace

In general, the light source for a scene will not be white. (A white source is one that emits the same flux of photons at each wavelength.) Suppose we take two spectral samples of image intensity S and L. Spectral normalization is any procedure that transforms S and L into new values S^* and L^*, where the latter measurements *would have been obtained* had the illuminant been white.

Land's normalization is $(S^*, L^*) = (\frac{S}{S_{max}}, \frac{L}{L_{max}})$, where S_{max} and L_{max} are the greatest intensities measured in the S and L samples throughout the image.

Our theory of normalization is based on the spectral crosspoint, as discussed in section 4. To relate our current work to Land's experiments, we must ask how spectral crosspoints appear in Landspace. We claim that a crosspoint corresponds to a pair of points in Landspace that form a line segment of negative slope (in Landspace). To avoid confusion, we will refer to the slope of line segments in Landspace as "Landslope," as distinguished from *spectral slope* in plots of intensity versus wavelength as discussed earlier in the paper. (Landslope, then, is a function of a pair of regions, whereas spectral slope is a property of a single region.) Our claim, again, is that a spectral crosspoint corresponds in Landspace to a pair of points of negative Landslope. The proof follows.

Suppose there is a crosspoint between regions X and Y. Then, say, $S_X > S_Y$ and $L_X < L_Y$. Does the crosspoint imply some sort of relationship among the Landspace coordinates for X and Y, (S_X^*, L_X^*) and (S_Y^*, L_Y^*)? It's easy to see from the definition of Landspace coordinates that $S_X^* > S_Y^*$ and $L_X^* < L_Y^*$. Now Landslope is given by $\frac{L_X^* - L_Y^*}{S_X^* - S_Y^*}$, so the Landslope of crosspoint regions X and Y is negative. (Note the assignment of S^* to the abscissa is irrelevant to the result.)

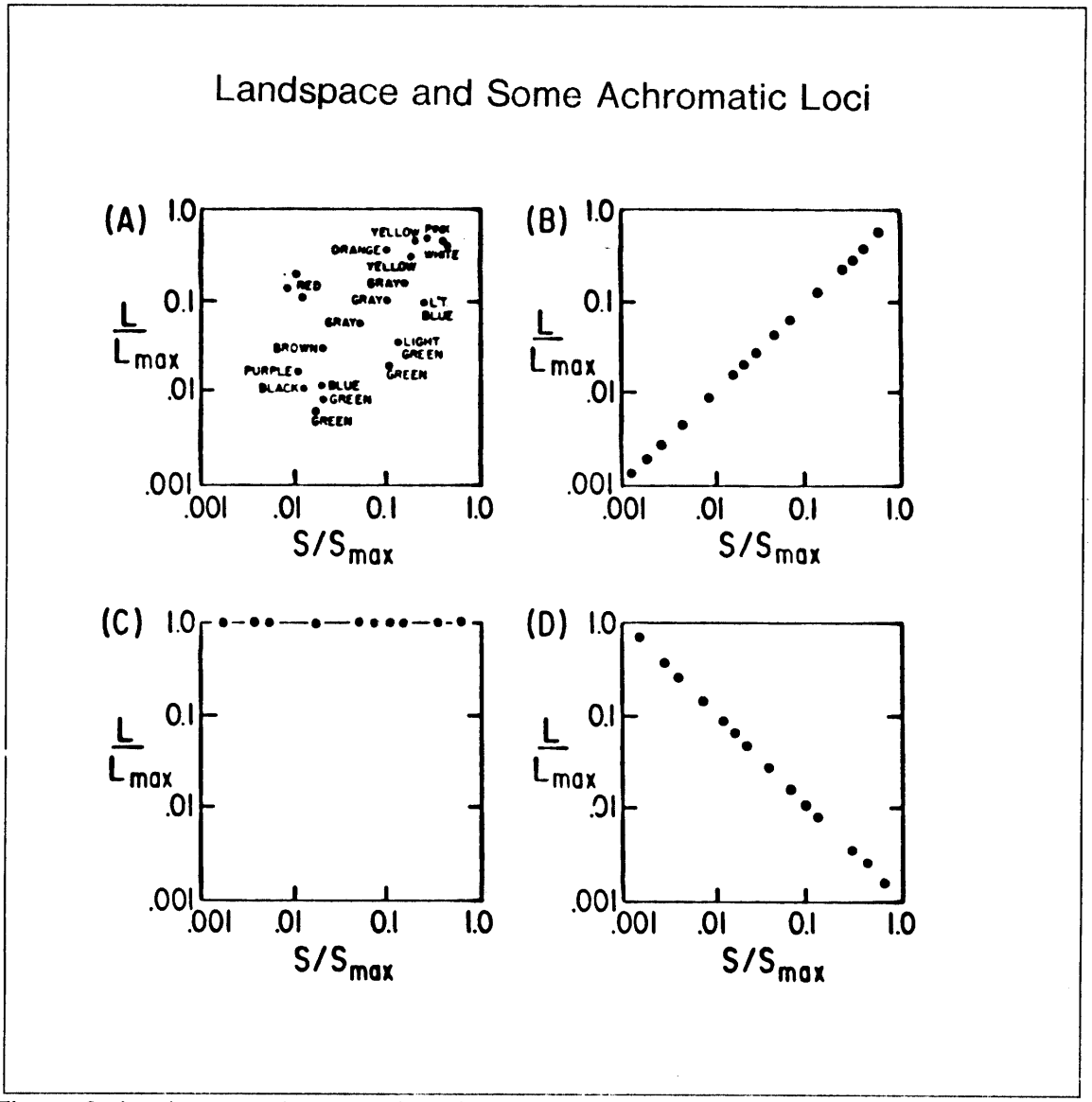

Figure 6 Landspace and some of its achromatic loci, as discovered by Land (1959a). A) Land's coordinate system (adapted from Fig. 1 of Land, 1959b) that relates perceived hue to the fraction of maximum long- and short-wavelength light (expressed on log-log axes). This coordinate system we call "Landspace." B) Image regions correspond to a line of unit Landslope. Such an image (as well as the next two) results in a monochromatic percept. (This is produced by placing identical records in the long- and short-wavelength projectors.) C) A line of zero slope. (This is created by removing the record from the long-wavelength projector.) D) A line of slope −1. (One record is placed in the short-wavelength projector, and its photographic negative is placed in the long-wavelength projector.)

7.2.3 Failures of Landspace

Landspace is a way of predicting the perceived hue of a region given the ratio of its intensity to the maximum intensity, at long and short wavelengths. This predictive scheme is successful for two-color projection of natural images. Land noticed, however, that for certain *contrived* images, his coordinate system failed totally. These images were seen as achromatic (or monochromatic). What did these concocted failure conditions have in

common? Land (1959a) suspected that "any arrangement which yielded points falling on a straight line [in Landspace], or even on a simple smooth curve, would be colorless." [Judd (1960) formalized Land's results on failure conditions.] We will show that the failure conditions Land has discovered correspond to situations in which our theory is unable to make any material distinctions. Furthermore, we will show our theory predicts stricter failure conditions than does Land in his conjecture.[16]

Figures 8b,c,d depict three concocted situations that Land found [and Judd (1960) verified] to cause a breakdown of Landspace. In figures 6b and 6c, the failure loci are straight lines of non-negative Landslope. Notice that for such loci, *there can be no spectral crosspoints,* since crosspoints correspond to point-pairs of negative Landslope.

Our normalization scheme, re-cast in Landspace, calls for inspection of all point-pairs of negative Landslope, since this subset of points is more likely to arise from a random set of materials than the totality of points. So a visual system using our normalization procedure, finding no point-pairs of negative Landslope (no crosspoints), would fairly conclude that there are no material changes and hence only a single material is present. A monochromatic (or achromatic) percept is an apt result, then, for a system that encodes material type by hue.

Consider next a collinear collection of points of negative Landslope in Landspace. Normalization can proceed according to our scheme, since spectral crosspoints are available. Thus we disagree with Land's (1959a) conjecture that *all* collinear sets of points will be failures. We predict the locus shown in Fig. 7a will yield a range of hues. Only collinear sets of positive Landslope will fail.

There is one special exception to our prediction that collinear loci of negative Landslope will produce chromatic percepts. A set of points of Landslope -1 (Fig. 6d) corresponds to an isoluminance image. Such an image has no luminance edges, and has long been known to disrupt vision (Evans, 1948). We have argued elsewhere (Rubin & Richards, 1982) that crosspoints are only meaningful across edges, and hence should only be sought across luminance discontinuities. Thus the isoluminance condition (the locus of Landslope -1) implies an absence of crosspoints and a failure of normalization, leading to an achromatic percept.

We turn next to Land's conjecture that curved loci in Landspace will yield achromatic percepts.[17] We believe this is an overgeneralization. We predict, along with Land, that the

[16]Land's conjecture (that smooth one-dimensional loci in Landspace will be seen as achromatic) is problematic. It seems difficult to legislate whether a collection of points in a plane constitutes a curvilinear arrangement or defines an area. A smooth curve can be drawn through *any* collection of points in a plane.

[17]A line in Landspace does not have absolute significance anyway, since linehood depends on the choice of axes. For example, a line in Landspace with log-log axes will not be a line with linear or power axes.

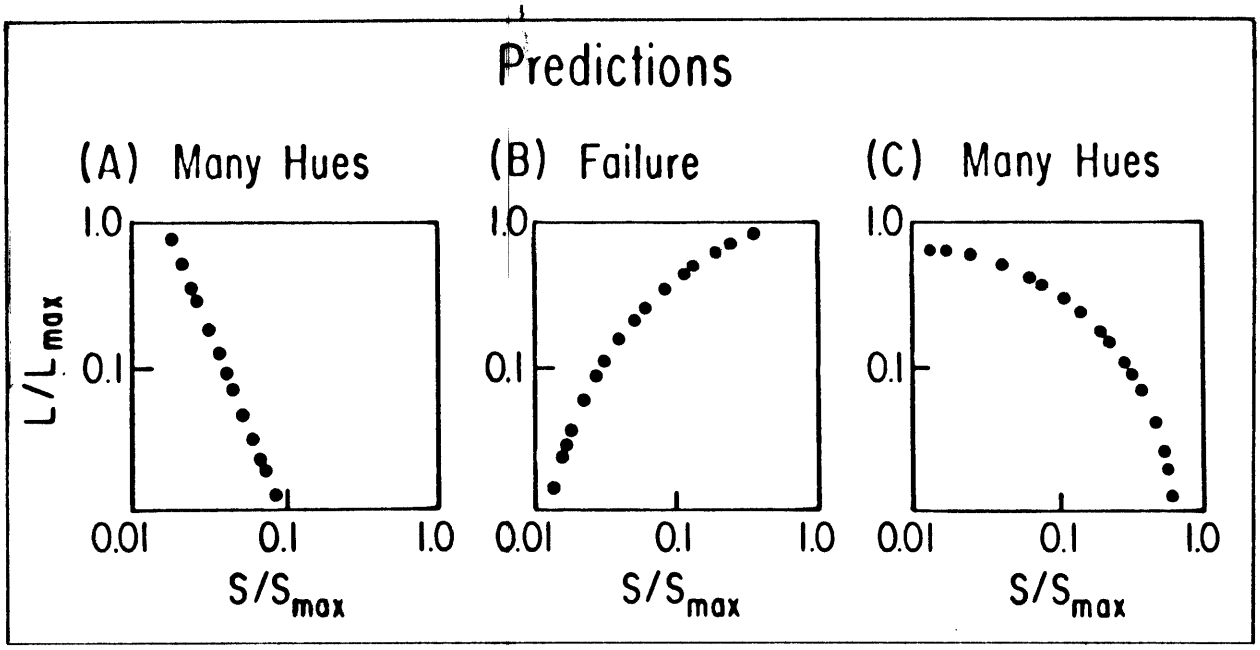

Figure 7 Predictions of our theory that conflict with Land's conjecture that all one-dimensional loci in Landspace will yield achromatic (monochromatic) percepts. A) A linear locus of negative Landslope ($\neq -1$). B) A smooth locus of points without point-pairs of negative Landslope should be a failure condition. C) A smooth locus of points that has point-pairs of negative Landslope should produce a range of hues.

non-linear locus of points in Landspace shown in Fig. 7b, since it contains no point-pairs of negative Landslope (no crosspoints), will be achromatic. In contrast, the one-dimensional locus of Fig. 7c has point-pairs of negative Landslope, and should yield a range of hues.

To sum up, we have suggested that failure conditions of two-color projection occur when there are no spectral crosspoints. That is, Land failures should occur if there are no (or too few) point-pairs of negative Landslope. Our predicted range of failure conditions is therefore narrower than Land's. Furthermore, Land's account of failures is purely descriptive; ours is explanatory (via the theory of material changes).

7.2.4 Opposite Slope Sign and Landspace

We have argued that the opposite slope sign condition (between two regions) is strong grounds for inferring the two regions are composed of different materials. Can this condition be recast in Landspace? The answer is yes: *two regions in the opposite slope sign relation map to two points straddling the line of unit Landslope in Landspace.*[18]

The argument is as follows. Image regions X and Y satisfy the opposite slope sign

[18]Land's early work relied on two spectral samples. Thus there is only one opposite slope sign condition to worry about, as shown in Fig. 5. Our trichromatic theory, sketched in Fig. 6, is not applicable to Land's work.

condition if normalized image intensities[19] obey the following: $L_X^* > S_X^*$ and $L_Y^* < S_Y^*$, where L_X^* denotes the normalized intensity in the longwave sample of region X, and so on. But the last condition indicates that (S_X^*, L_X^*) lies above the line of unit Landslope in Landspace (given the abscissa marks S^* values), and (S_Y^*, L_Y^*) lies below. If the Land normalization is correct, then we have shown that two regions in an opposite slope sign condition map to a pair of points in Landspace straddling the line of unit Landslope. (For many complex natural images, Land's normalization scheme and ours could yield similar results. That is why we can accept Land's scheme as approximately correct.)

Examine again Land's results shown in Fig. 6a. Land observed that the hues appearing above the line of unit Landslope are all "warm," and those falling below are "cool." (Wilson & Brocklebank [1960], in a study of two-color projection phenomena, noted that although hue, saturation, and lightness were not precisely preserved in the two-color reconstruction of the original still-life, at least the warm/cool aspect of hue was invariant.) The distinction between warm and cool colors is certainly the most fundamental fact of categorical hue perception. To sum up, given that Land's normalization has been successful, different materials (as discovered by the opposite slope sign criterion) map in Landspace to two points straddling the line of unit Landslope (and vice versa). In turn, two points straddling the unit slope line correspond to two qualitatively distinct hues, one warm and one cool. This observation supports our claim that hue is encoding information about differences in material.

7.2.5 Doubling the Record

Land discovered that if he modified the two-color paradigm by placing a second long-wave record, say, in the long-wave projector, perception is not substantially changed. How does this transformation alter Landspace coordinates? The longwave coordinates of Landspace are squared (and hence reduced since Landspace coordinates are between zero and unity). The shortwave coordinates are unchanged.

Notice that to the extent that this "doubling the record" manipulation leaves perceived hue unchanged, Landspace has failed. Landspace was intended to allow predictions of perceived hue for a given pair of filters. But a successful prediction for the normal two-color set-up will be unsuccessful for the doubled record. For example, in the normal set-up, the line of greys is the line of unit slope; when the record is doubled and perception remains the same, the line of greys shifts to some curved locus (or a line of Landslope two on log-log axes). The perception of hue therefore depends not only on the two filters, but *the distribution of values in Landspace* as well.

[19]Recall that in order for the opposite slope sign to reliably indicate material changes, the illuminant must be white, or equivalently, the image must be normalized.

How does the doubling-the-record transformation affect our normalization scheme? Crosspoints are preserved, since doubling the record preserves ordinal relations among the S measurements and among the L.

Recall that the premise of our normalization theory was that given a random collection of materials under white light, the median spectral slope would be zero. Suppose we divided the random collection into two subgroups, one of light materials (high average albedo over wavelength) and one of dark. It seems that our fundamental premise would also be true of each of the subgroups. That is, it seems reasonable to guess that the median spectral slope of a random assortment of dark materials would be zero, and similarly for the light collection.[20] We suggest that our normalization could proceed independently for several subgroups of image intensities, all members of a subgroup sharing similar lightness values. Such an extension of our normalization scheme would tolerate some "bending" of the grey locus, as occurs when a record is doubled.

It's worth asking whether any natural situations cry out for normalization that varies with intensity, such as we propose. That is, where might we expect the spectral character of illumination to vary with intensity? Aquatic environments come to mind. At a given depth, light coming from below has lower intensity and lower peak wavelength than does light coming from above (Levine & MacNichol, 1982; McFarland & Munz, 1975a). Two distinct Landspaces could be set up, one for the lower visual hemifield, another for the upper. In terrestrial habitats, it might be desirable to normalize shadowed (low intensity) regions separately from fully lit (high intensity) regions in the same image. This would be the case if the diffuse light were bluer than the direct light due to scatter.[21]

7.3 Neurophysiological Operators

How might operators be constructed that would detect crosspoints and the opposite slope sign condition? Consider crosspoints first. Suppose we would like to detect crosspoints of the sort shown in Fig. 8a. Two spectral samples, M and L, are taken on both sides X and Y of an edge. Figure 8b shows schematically such a detector. Intensity values of the L samples are compared across the edge. We define $\binom{L^+}{L^-}$ as the difference between the L value of side Y from the L value of side X, if that value is positive, and zero otherwise. Similarly, the M samples are compared across the edge. Here the M value of side X is subtracted from that of side Y (again with a zero minimum). The results of the

[20]Of course, this assumption can be tested. If dark materials, say, tended to be reddish, our assumption would be incorrect and require modification.

[21]Note that in this case, the grey world assumption of section 3.4 is inappropriate.

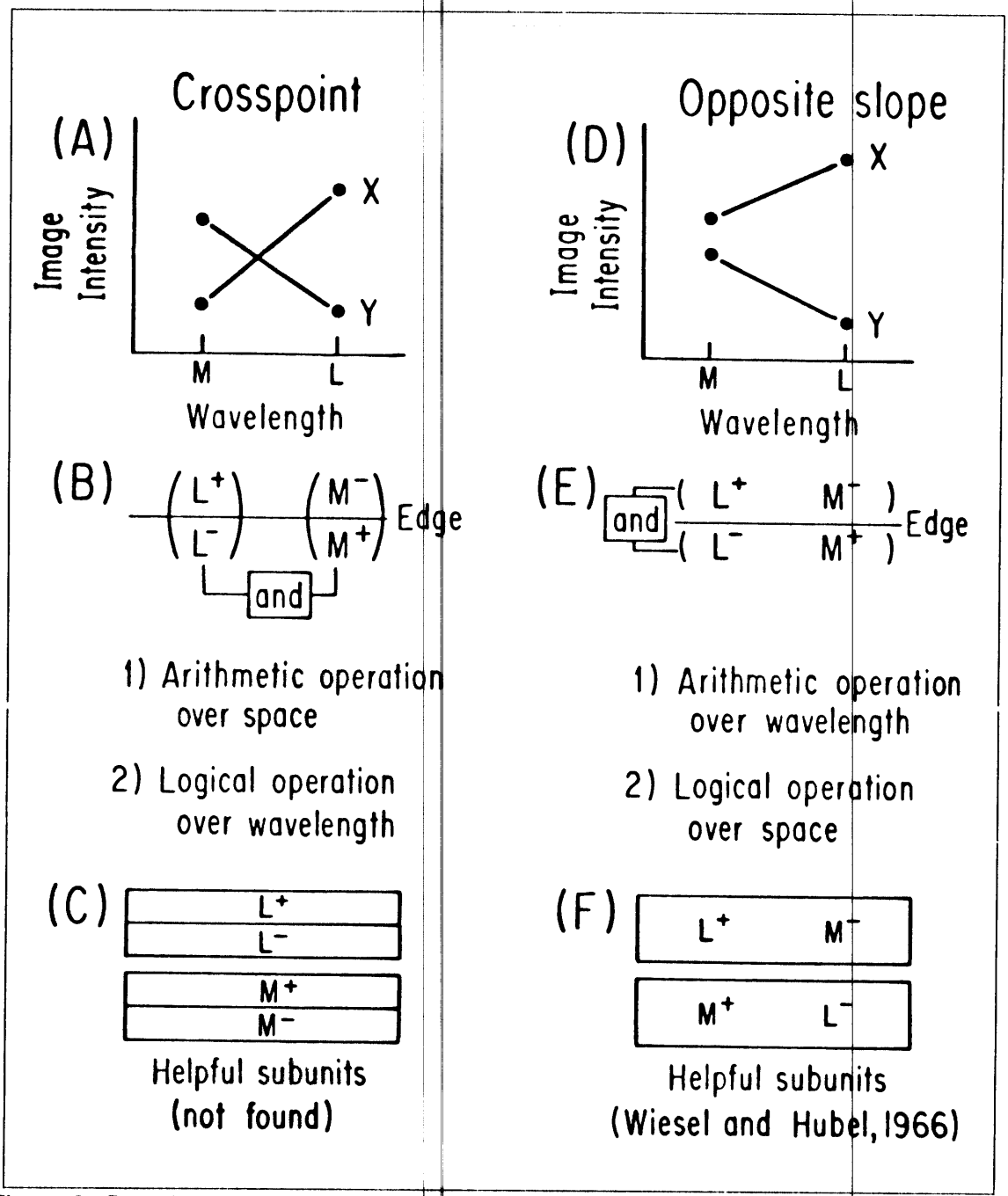

Figure 8 Detecting crosspoints and the opposite slope sign condition. a) The spectral crosspoint to be detected. b) A schematic crosspoint detector. Note a logical spectral operation follows an arithmetic spatial operation. c) Useful intermediate units for crosspoint detection are sketched. Split bar-shaped configuration indicates a spatial comparison. Comparisons are made within spectral channels. Outputs of these intermediates can then be combined logically. d) The opposite slope sign condition to be detected. e) A schematic detector for the opposite slope sign. Note a logical spatial operation follows an arithmetic spectral operation. f) Useful intermediate units for opposite slope sign detection. Spectral comparisons are made in a single spatial region. Outputs of these intermediates can then be combined logically.

two *arithmetical* operations $\binom{L^+}{L^-}$ and $\binom{M^-}{M^+}$ are combined *logically* with an **AND**.[22]

Next consider the detection of opposite slope signs as shown in Fig. 8d. On each side of the edge, a spectral comparison must be made. The arithmetic operation $(L^+ \quad M^-)$ denotes the value of the M sample subtracted from that of the L sample on one side of the edge. This operation has a minimum value of zero. The spectral comparison $(L^+ \quad M^-)$ is then **AND**ed with the similarly defined comparison $(L^- \quad M^+)$ on the opposite of the edge to create the opposite slope sign detector shown in Fig. 8e.

It's worth noting that for the crosspoint detector, two spatial, arithmetic operations are logically combined across wavelength samples. For the opposite slope sign detector, two spectral, arithmetic operations are combined logically across space (across an edge).

Both detectors (Figs. 8b and 8e) superficially resemble double-opponent units described in many species (Daw, 1972; Michael, 1978a; Livingstone & Hubel, 1984). That is, both detectors have two spatial fields—one that receives excitatory connections from one spectral sample, and inhibitory inputs from another spectral sample, and a second spatial field receiving the opposite spectral inputs. Neurophysiological tests have not yet been sufficiently detailed to distinguish whether double opponent units are computing spectral crosspoints, or opposite slope signs, or neither.

Useful intermediate units for crosspoint detectors are shown in Fig. 8c. The line through the bar-shaped configuration indicates a comparison across space. Comparisons are within spectral channels; outputs of these intermediates can be combined logically to build crosspoint detectors. This sort of intermediate unit—spatial comparisons in isolated spectral channels—has only been reported once (Michael, 1978b), apparently without replication.

Figure 8f depicts useful intermediate units for opposite slope sign detection. Spectral comparisons are made within a single spatial region. Units of this type—spectrally opponent but spatially undifferentiated—have been described by physiologists recording from monkey lateral geniculate nucleus (Weisel & Hubel, 1966; Krüger, 1977; Michael, 1978b). This evidence suggests that the double-opponent units described in primate V1 might be performing an opposite slope sign computation. Detailed color neurophysiology is needed to test this notion.

[22]Strictly speaking, the results of the two arithmetical operations must be converted to 0 or 1 before being logically combined. Alternatively, the results of the arithmetical operations—the modified subtractions—could be multiplied together. A non-zero product implies a crosspoint.

8.0 Summary

Our theory of color vision presents two types of operators—the spectral crosspoint for normalization and the opposite slope sign—which suffice in most cases to normalize for the illuminant and to categorize the albedoes in the scene. Our scheme should differentiate between the common natural pigments (chlorophylls, xanthophylls and flavanoids), for example, but not between the density variations of any one of these pigments. The theory does not address this latter problem—namely how we appreciate the fine changes in the grain of a piece of teakwood. A quantitative color vision system, of greater complexity than the qualitative computations described here, will be needed for such fine discriminations. Categorical color vision is simply an inexpensive method for making rapid and reliable coarse judgments about materials.

Appendix 1: Lawful Processes

This appendix shows that image edges that arise from 1) change in surface orientation, 2) pigment density variations, 3) shadows and 4) highlights all preserve the ordinal relations of image intensities across that edge, and hence cannot cause the opposite slope sign condition.

1.0 Surface Orientation Change

Let X and Y be regions on either side of an edge due solely to a surface orientation discontinuity. Then the image intensities (as functions of wavelength) $I_X(\lambda)$ and $I_Y(\lambda)$, measured in X and Y, respectively, are related multiplicatively. That is, $I_X(\lambda) = \alpha I_Y(\lambda)$ for some constant α (Rubin & Richards, 1982; Horn & Sjoberg, 1979). Two functions differing only by a multiplicative constant have identical ordinality.

2.0 Pigment Density Variation

Suppose X and Y are two regions on a planar piece of a single material that differ only in pigment density. Then if the albedo (as a function of wavelength) of region X is $\rho(\lambda)$, the albedo of Y can be approximated[23] by $\rho^b(\lambda)$, where b is a constant related to pigment density (Rubin & Richards, 1982; Wyszecki & Stiles, 1967).

The light measured from regions X and Y is the product of the albedoes of X and Y with the radiant intensity of the illuminant. Since X and Y are assumed coplanar (recall that pigment density change is stipulated as the *sole* cause of the edge), and the illumination is the same for both, then any difference between measured intensities from the two regions will be due to a difference in albedo functions. But the albedo functions are related by an exponential constant, and two functions so related have identical ordinality. Therefore, image intensities across a pigment density change will have identical ordinality. [Examples of this relation for natural pigments can be seen in Krinov (1971), Francis & Clydesdale (1975) or Snodderly (1979).]

3.0 Shadow

Consider an edge separating a lit region from a shaded one. Both lit and shaded regions reflect diffuse illumination toward the viewer. The lit region, in addition, reflects a

[23] This exponential relation presumes that the embedding material is spectrally neutral. If the embedding layer reflects different wavelengths unequally, then change in pigment density has a more complex description. In particular, pigment density changes can mimic material changes.

direct source. If $I_{lit}(\lambda)$ and $I_{shad}(\lambda)$ are image intensities (as functions of wavelength λ) from lit and shaded regions, respectively, then:

$$\begin{aligned} I_{shad}(\lambda) &= E_{diffuse}(\lambda)\rho(\lambda) \\ I_{lit}(\lambda) &= [E_{diffuse}(\lambda) + E_{direct}(\lambda)]\rho(\lambda) \end{aligned} \tag{1}$$

where $E_{diffuse}(\lambda)$ and $E_{direct}(\lambda)$ are the diffuse and direct components of illumination, and $\rho(\lambda)$ characterizes the albedo of the material.

By inspection of equations (1), it is clear that ordinality can be violated in the case of shadow. That is, a false target is possible. The visual world, fortunately, offers certain regularities. There is usually some close relation between diffuse and direct illumination (Goral *et al.*, 1984). This is not surprising, since diffuse light results from diverse, random reflections of the direct light from a variety of materials in the scene. An assumption will be made that this is usually the case: a visual system can presume that diffuse light has the same spectral character as the direct light. That is, $E_{diffuse}(\lambda) = kE_{direct}(\lambda)$, for some constant k. This we call the "grey world" assumption (see Section 3.1), because it is implied by the statement that all the albedoes of a scene will average to grey. Anecdotal data support the grey world assumption. Hailman (1979) measured spectral irradiance functions in a pine woods in a sunny area and in nearby shade. The functions are strikingly similar in shape, and are shown in Fig. 9.

Invoking the grey world assumption, equations (1) become:

$$\begin{aligned} I_{shad}(\lambda) &= kE_{direct}(\lambda)\rho(\lambda) \\ I_{lit}(\lambda) &= (1 + k)E_{direct}(\lambda)\rho(\lambda) \end{aligned} \tag{2}$$

Note that the lit and shaded regions now give rise to multiplicatively related image intensity functions. Ordinality will therefore be preserved.

4.0 Highlights

The analysis of highlights is slightly more complex. The following equations (Rubin & Richards, 1982; equations 14a) express the image intensities to be found in a highlight and neighboring matte region:

$$\begin{aligned} I_{matte}(\lambda) &= (E_{diffuse}(\lambda) + E_{direct}(\lambda))\rho(\lambda) \\ I_{highlight}(\lambda) &= \delta E_{direct}(\lambda) + (1 - \delta)[E_{diffuse}(\lambda) + E_{direct}(\lambda)]\rho(\lambda) \end{aligned} \tag{3}$$

where $I_{matte}(\lambda)$ and $I_{highlight}(\lambda)$ are the images intensities (as functions of wavelength) in matte and highlighted regions, and $\delta \in (0, 1)$ is a constant that indicates to what extent the surface is mirrorlike ($\delta = 1$ describes a perfect mirror). (See Richards, Rubin & Hoffman, 1982, for a more extended treatment.)

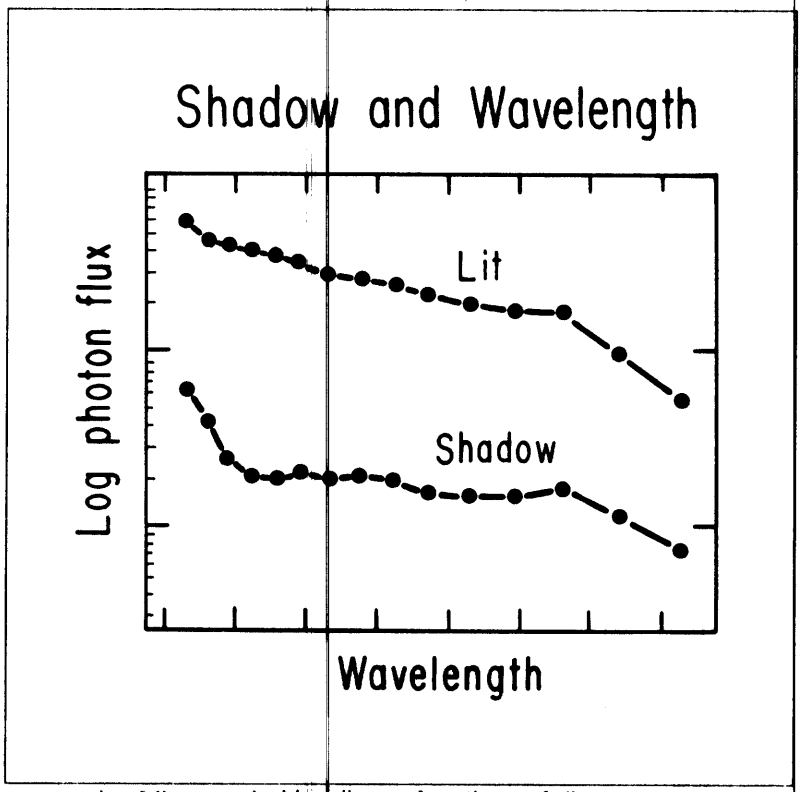

Figure 9 Measurements of the spectral irradiance functions of direct sunlight and nearby shade in a Florida pine woods, adapted from Hailman (1979), Fig. 74a. On the ordinate is the logarithm of photon flux. The abscissa shows wavelength.

The equations express the fact that both highlighted and matte regions reflect both direct and diffuse light. In addition, the highlight, acting as a partial mirror, reflects the direct light.

Applying the grey world assumption, equations (3) become:

$$\begin{aligned} I_{matte}(\lambda) &= (1+k)E_{direct}(\lambda)\rho(\lambda) \\ I_{highlight}(\lambda) &= \delta E_{direct}(\lambda) + (1-\delta)(1+k)E_{direct}(\lambda)\rho(\lambda) \end{aligned} \tag{4}$$

which reduces to

$$\begin{aligned} I_{matte}(\lambda) &= (1+k)E_{direct}(\lambda)\rho(\lambda) \\ I_{highlight}(\lambda) &= E_{direct}(\lambda)[\delta + (1-\delta)(1+k)\rho(\lambda)] \end{aligned} \tag{5}$$

By inspecting equations (5), it can be seen that highlights can produce a spurious violation in ordinality. Assume now that the image has been normalized with respect to the color of the illuminant. Normalization is any scheme that allows recovery of the spectral character of the illuminant. (Such a computation is presented in section 4.) Normalization is equivalent to a transformation of the image intensities to what they would have been had the illuminant been white; it allows us to set $E_{direct}(\lambda) = \beta$, where β is some constant.

Both equations (5) can now be rewritten substituting β for $E_{direct}(\lambda)$, yielding

$$I_{matte}(\lambda) = \beta(1 + k)\rho(\lambda)$$
$$I_{highlight}(\lambda) = \beta[\delta + (1 - \delta)(1 + k)\rho(\lambda)] \quad (6)$$

With the two assumptions of grey world and spectral normalization, highlights will not produce violations in ordinality. This can be seen in equations (6), where the image intensity function of the highlighted region is simply related to the image intensity function of the neighboring matte region. The intensity in the matte region is multiplied by a constant $(1 - \delta)$, and then a constant function $(I(\lambda) = \delta\beta)$ is added. These two operations preserve ordinality; hence no opposite slopes will arise given our assumptions.

Appendix 2: Algorithm for Spectral Normalization and Material Categorization

Given a full-color image of a scene lit by an unknown illuminant, and a way of finding edges and regions, regions can be assigned to one of a small number of material categories. Regions in different categories are made of different materials. An algorithm for categorizing materials is sketched below. The first step is to correct for colored illumination; the second is to categorize.

In the Beginning

The original full-color image can be viewed through three spectral filters, yielding three distinct maps of image intensity, say R, G, and B. See Fig. 10a. These three maps of image intensity we call "spectral images." The number of filters, or their spectral characteristics, should not be important. All that matters is that the filters yield independent measurements.

Spectral Normalization

First, apply an edge operator to the image. The particular edge operator should not be crucial. Assume the edge operator produces a closed set of edges.[24] Next, edge segments must be made explicit. See Fig. 10b. This involves understanding vertices. For example, a T-vertex terminates the edge that is the leg, but not the edge that's the crossbar. Identifying edge segments is important because we will iterate through a list of them.

For each edge segment, two narrow strips must be defined, one on each side. Call the strips X and Y. (Understanding vertices is important because the strips must be free of edges.) See Fig. 10c.

Average the intensity values of each of the spectral images R, G, and B in both the X and Y strips. The output of this step is six values R_X, R_Y, G_X, G_Y, B_X, and B_Y.

For each edge segment, check for two types of crosspoint, RG, and BG.[25] (The conditions are $(R_X - R_Y)(G_X - G_Y) < 0$ and $(G_X - G_Y)(B_X - B_Y) < 0$, respectively.) Note the possibility of a third crosspoint involving the R and B samples.

Suppose an image has n crosspoint edge segments. For each crosspoint, record

[24] If algorithm for edge detection does not produce closed edges, then regions must somehow be identified using edge fragments.

[25] The R and G samples can yield crosspoints, and independently, so can the B and G samples. The G sample could just as easily be taken as the photopic luminosity function.

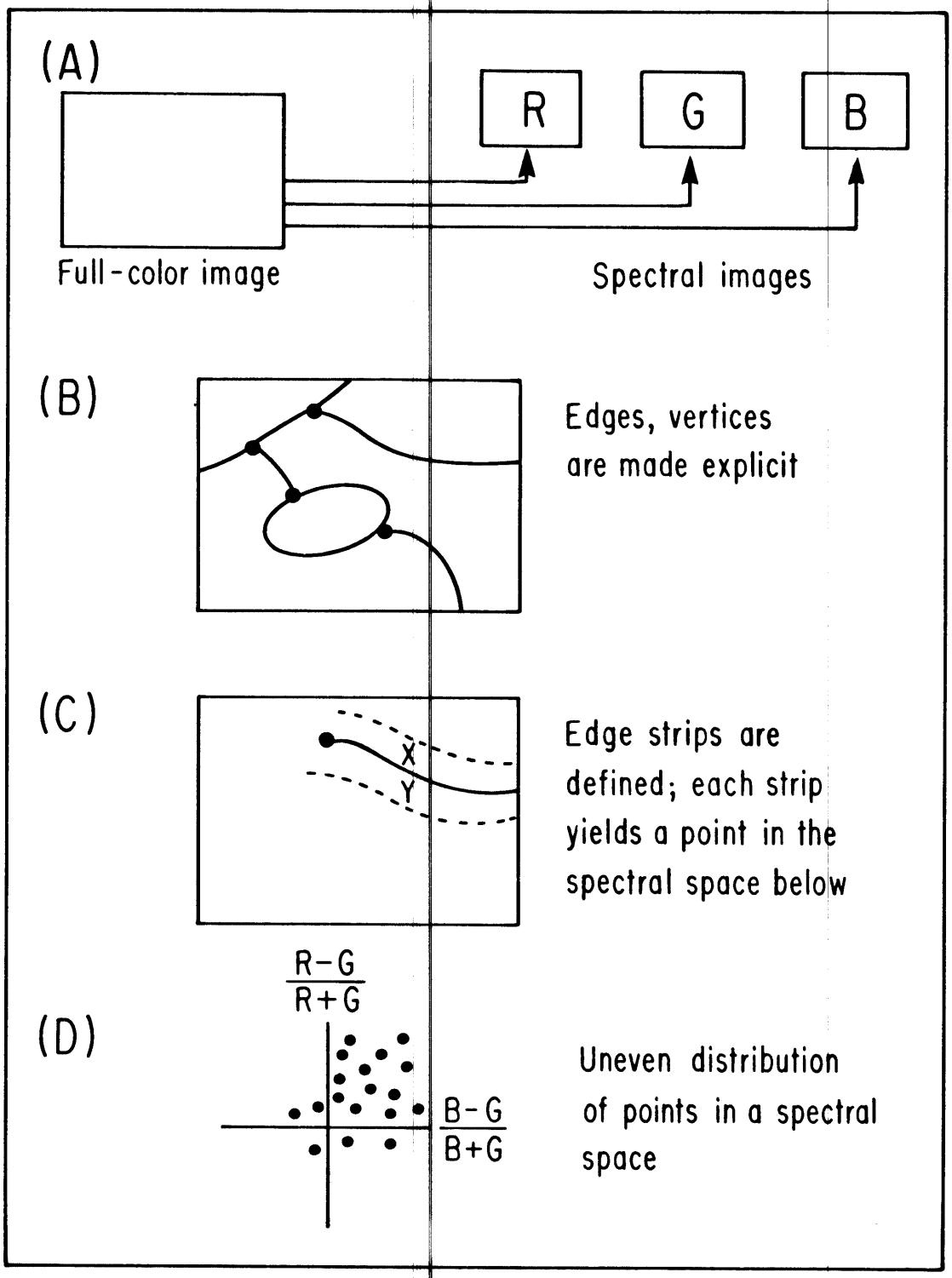

Figure 10 a) The full-color image is run through three spectral filters R, G, and B. b) Edge segments have been found and made explicit. This image shows five edge segments. Vertices have been found, and are here marked with large black dots. c) On either side of one of the edges, narrow strips X and Y are defined. No edge segments should be in the strips. Intensity averages will be taken in the three spectral images in both of the strips, yielding six measurements. This is done for each edge segment in the image. d) Measurements taken from strips about each edge map to points in a spectral space defined by axes as labeled. Normalization consists of multiplying R and B values by factors such that equal numbers of points will be found in each quadrant.

spectral information about the two abutting strips. In particular, store two color contrast values per region:

$$\frac{R_i - G_i}{R_i + G_i}, \frac{B_i - G_i}{B_i + G_i}, \quad i = 1, \ldots, 2n \tag{8}$$

where i is an index ranging over the $2n$ edge strips defined around n crosspoints. This particular form of ratio is useful because its value must lie in the closed interval $[-1, 1]$. The spectral information recorded can be considered as $2n$ points in a two-dimensional spectral space (with axes of $\frac{R-G}{R+G}$ and $\frac{B-G}{B+G}$) shown in Fig. 10d. (See also Fig. 4a.)

Let $\mathcal{U}$ be the number of points in the upper half-plane of the spectral space (Fig. 10d), and $\mathcal{L}$ be the number of points in the left half-plane. Under a white illuminant, we'd expect a random assortment of materials to yield $\mathcal{U} \approx \mathcal{L} \approx n$; that is, points should be roughly equally distributed among the quadrants of the spectral space.

If the $2n$ points are not divided equally among the quadrants of the spectral space, we must seek normalization constants α and β that satisfy the following criterion:

$$MEDIAN\left[\frac{\alpha R_i - G_i}{\alpha R_i + G_i}\right]_{i=1,\ldots,2n} = MEDIAN\left[\frac{\beta B_i - G_i}{\beta B_i + G_i}\right]_{i=1,\ldots,2n} = 0 \tag{9}$$

For a large enough number of image regions, we can take

$$\alpha = \frac{1 - \overline{C}_{RG}}{1 + \overline{C}_{RG}} \qquad \beta = \frac{1 - \overline{C}_{BG}}{1 + \overline{C}_{BG}} \tag{10}$$

where $\overline{C}_{RG}$ and $\overline{C}_{BG}$ are means of the sets of measurements (8):

$$\overline{C}_{RG} = \frac{1}{2n}\sum_{i=1}^{2n}\frac{R_i - G_i}{R_i + G_i} \qquad \overline{C}_{BG} = \frac{1}{2n}\sum_{i=1}^{2n}\frac{B_i - G_i}{B_i + G_i} \tag{11}$$

The values of α and β in (10) will provide a correct normalization (i.e., normalization criterion (9) will hold) given some simple statistical conditions.[26]

The correctness of the normalization constants α and β can easily be checked by verifying that criterion (9) holds. If not, the values of α and β can be adjusted incrementally in an iterative procedure. The entire normalization algorithm is shown as a flowchart in Fig. 11.

Once correct values of the normalization constants are returned by the algorithm, the three spectral images R, G and B can be transformed into a set of normalized spectral

[26]There must be at least 12 independent crosspoint edges, and the mean and median of the set of measurements $\{\frac{R_i-G_i}{R_i+G_i}\}$ must approach the same value as $i \mapsto \infty$, and similarly for the set of measurements $\{\frac{B_i-G_i}{B_i+G_i}\}$. (See Siegel, 1956.)

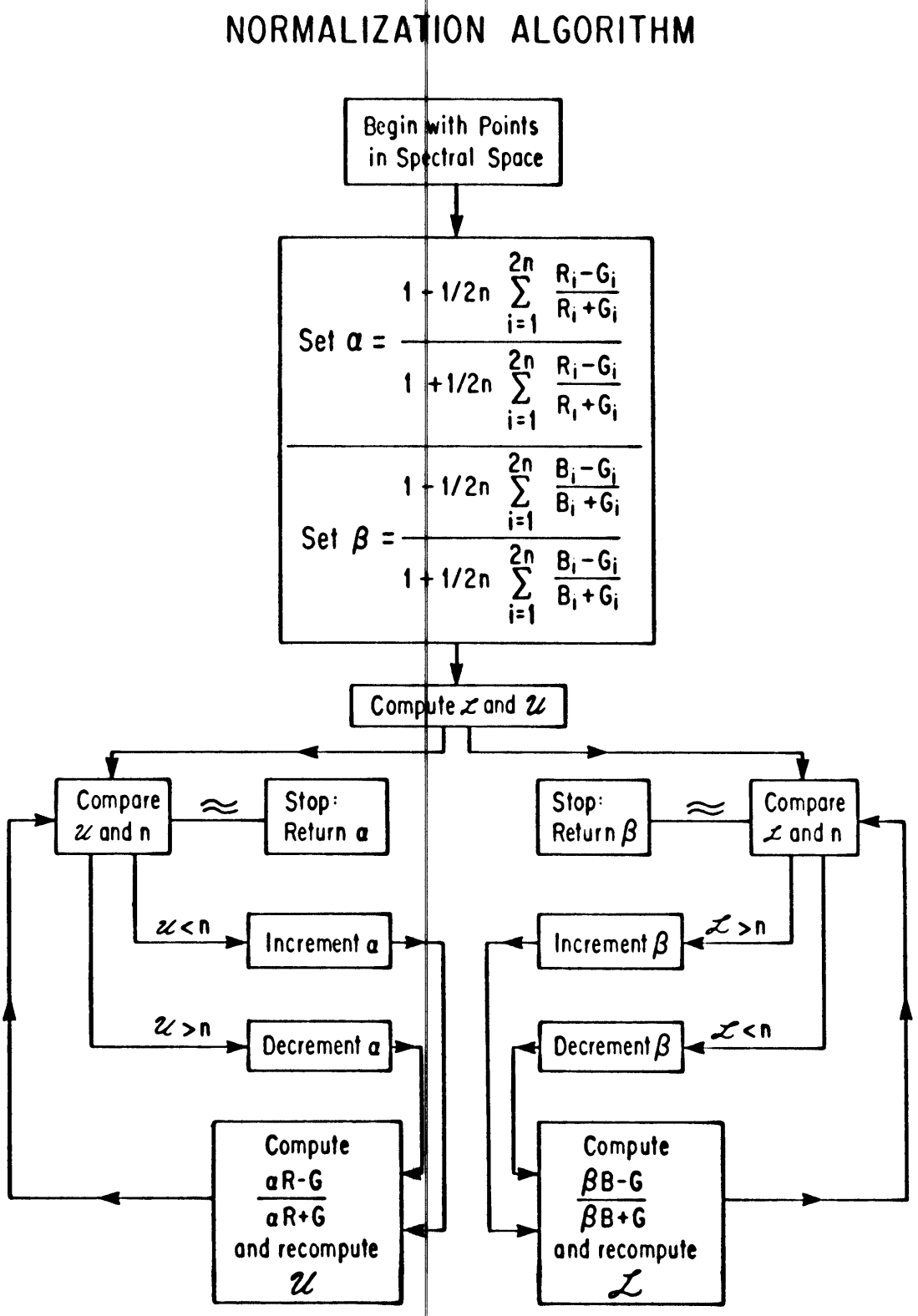

Figure 11 Normalization Flowchart. Begin with points scattered in spectral space, and end with a pair of multiplicative normalization coefficients, α to balance the R image with respect to G, and β to balance the B image with respect to G.

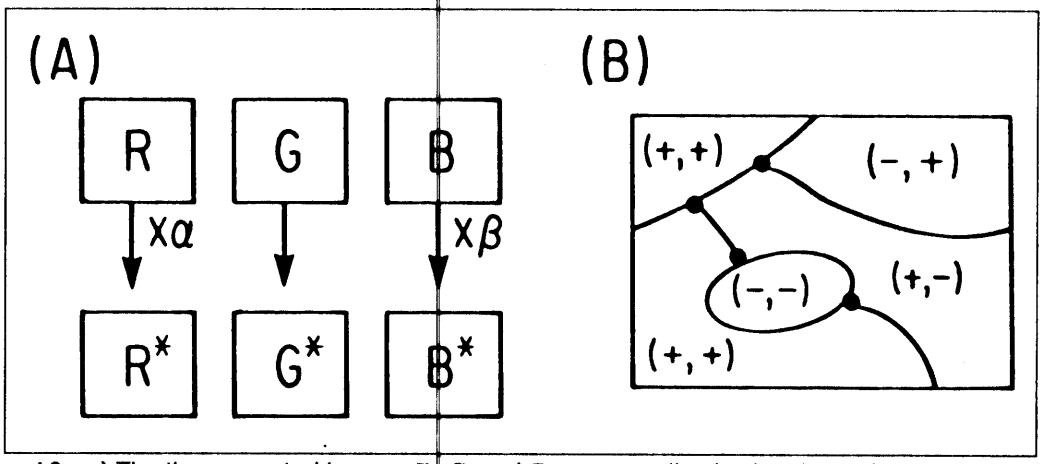

Figure 12 a) The three spectral images R, G, and B are normalized using the multiplicative constants produced by the procedure shown in Fig. 11. The normalized spectral intensity maps are R^*, G^*, and B^*. b) The regions of the image sketched in Fig. 10b labeled with material categories. Each region is assigned one of four possible ordinal doublets.

images. All values in the R image are multiplied by α, yielding R^*. (The asterisk superscript denotes normalized intensity; see section 7.2.2.) Similarly, $B^* = \beta B$. Spectral image G is unchanged: $G = G^*$. See Fig. 12a.

Spectral Categories

Suppose that when closed edge segments were found that image regions were made explicit. For each region i, measure the average values of the normalized spectral images, yielding the triplet (R_i^*, G_i^*, B_i^*). A triplet of numbers yields one obvious pair of ordinal relations:

$$(R_i^*, G_i^*, B_i^*) \mapsto (sign_{RG}, sign_{BG})_i \tag{12}$$

where $sign_{RG}$ is "+" if $G_i^* > R_i^*$, and "−" otherwise.

Each region can therefore be assigned to one of four material categories: $(+,+)$, $(-,+)$, $(-,+)$, $(+,-)$. This is shown in Fig. 12b. Two regions that are in different categories are composed of distinct materials.

Note that a third ordinal relation is sometimes independent, the $R^* - B^*$ comparison. If this relation is included, six spectral categories obtain.

Finally, note that while the algorithm described here is categorical, continuous information has not been lost; it is still available for more refined purposes. For each region i, the continuous-valued coordinates

$$\left[\frac{R_i - G_i}{R_i + G_i}, \frac{B_i - G_i}{B_i + G_i}\right] \tag{13}$$

should be useful.

REFERENCES

Bornstein, M.H., Kessen, W. & Weiskopf, S., "Color vision and hue categorization in young human infants." *J. Exp. Psych: Human Perception and Performance,* **2**, 115–129, 1976.

Cohen, J., "Dependency of the spectral reflectance curves of the Munsell chips." *Psychon. Sci.,* **1**, 369–370, 1964.

Daw, N.W., "Color-coded cells in goldfish, cat, and rhesus monkey." *Invest. Ophthal.,* **11**, 411–417, 1972.

Evans, R.M., *An Introduction to Color.* Wiley, New York, 1948.

Francis, F.J. & Clydesdale, F.M. *Food Colorimetry: Theory and Applications.* AVI Publishing Co., Westport, Conn., 1975.

Goethe, J.B. von, *Zur Farbenlehre, Didaktischer Teil,* 1808. See R. Matthaei, ed., "Goethe's Color Theory", 1970. Van Nostrand Reinhold Co., New York.

Goral, G.M., Torrance, K.E., Greenberg, D.P. & Battaille, B., "Modelling the interaction of light between diffuse surfaces." *Computer Graphics* **18(3)**, 213–222, 1984.

Hailman, J.P., "Environmental light and conspicuous colors." Ch. 7 (289–357) in *The Behavioral Significance of Color,* E.H. Burtt, Jr., ed., Garland Press, New York, 1979.

Hering, E., *Grundzuge der Lehre vom Lichtsimm.* Julius Springer, Berlin, 1920.

Hering, E., *Zur Lehre vom Lichtsimm.* Carl Gerald's Sohn, Wien, 1878.

Hering, E. *Outline of a Theory of the Light Sense.* L.M. Hurvich and D. Jameson, translators. Harvard University Press, Cambridge, Mass., 1964.

Horn, B.K.P. & Sjoberg R.W., "Calculating the reflectance map." *Applied Optics* **18**, 1770–1779, 1979.

Judd, D.B., "Appraisal of Land's work on two-primary color projections." *J. Opt. Soc. Am.* **50**, 254–268, 1960.

Krinov, E.L., "Spectral reflectance properties of natural formations." National Research Council of Canada, Technical Translation 439, 1971.

Krüger, J., "Stimulus dependent color specificity of monkey lateral geniculate neurones." *Exp. Brain Res.* **30**, 297–311, 1977.

Land, E.H., "Experiments in color vision." *Sci. Am.* **200** (5), 84–99, 1959a.

Land, E.H., "Color vision and the natural image (Part II)," *Proc. Natl. Acad. Sci. USA,* **45**, 636–644, 1959b.

Land, E.H., "Recent advances in retinex theory and some implications for cortical computations: color vision and the natural image." *Proc. Natl. Acad. Sci. USA* **80**, 5163–5169, 1983.

Levine, J.S. & MacNichol, E.F., Jr., "Color vision in fishes." *Sci. Am.* **246 (2)**, 140–149, 1982.

Livingstone, M.S. & Hubel, D.H., "Anatomy and physiology of a color system in the primate visual cortex." *J. Neurosci.* **4**, 309–356, 1984.

Marr, D., *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman & Co., San Francisco, 1982.

McFarland, W.N. & Munz, F.W., "Part II: The photic environment of clear tropical seas during the day." *Vis. Res.* **15**, 1063–1070, 1975a.

McFarland, W.N. & Munz, F.W., "Part III: The evolution of photopic visual pigments in fishes." *Vis. Res.* **15**, 1071–1080, 1975b.

Michael, C.R., "Color vision mechanisms in monkey striate cortex: simple cells with dual-color receptive fields." *J. Neurophys.* **41**, 1233–1249, 1978a.

Richards, W.A., Rubin, J.M. & Hoffman, D.D. "Equation counting and the interpretation of sensing data." *Perception* **11**, 557–576, 1982.

Michael, C.R., "Color vision mechanisms in monkey striate cortex: dual-opponent cells with concentric receptive fields." *J. Neurophys.* **41**, 572–588, 1978b.

Rubin, J.M. & Richards, W.A., "Color vision and image intensities: when are changes material?" *Biol. Cybern.* **45**, 215–226, 1982.

Siegel, S., *Nonparametric Statistics.* McGraw Hill, New York, 1956.

Snodderly, D.M., "Visual discriminations encountered in food foraging by a neotropical primate: implications for the evolution of color vision." Ch. 6 (238–285) in *The Behavioral Significance of Color,* E.H. Burtt, Jr., ed., Garland Press, New York, 1979.

Wiesel, T.N. & Hubel, D.H., "Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey." *J. Neurophys.* **29**, 1115–1156, 1966.

Wilson, M.H. & Brocklebank, R.W., "Two–colour projection phenomena." *J. Photo. Sci.* **8**, 141–150, 1960.

Wright, A.A. & Cumming, W.W., "Color–naming functions for the pigeon." *Journal of the Exp. Anal. of Beh.*, **15**, 7–17, 1971.

Wyszecki, G. & Stiles, W.S., *Color Science: Concepts and Methods, Quantitative Data and Formulas.* Wiley, New York, 1967.