From *Computer Vision Systems*, A Hanson & E. Riseman (Eds.), pp. 3-26, New York: Academic Press (1978).

COMPUTER VISION SYSTEMS

RECOVERING INTRINSIC SCENE CHARACTERISTICS FROM IMAGES

H.G. Barrow and J.M. Tenenbaum,
SRI International,
Menlo Park, CA 94025.

ABSTRACT

We suggest that an appropriate role of early visual processing is to describe a scene in terms of intrinsic (vertical) characteristics -- such as range, orientation, reflectance, and incident illumination -- of the surface element visible at each point in the image. Support for this idea comes from three sources: the obvious utility of intrinsic characteristics for higher-level scene analysis; the apparent ability of humans to determine these characteristics, regardless of viewing conditions or familiarity with the scene; and a theoretical argument that such a description is obtainable, by a noncognitive and nonpurposive process, at least, for simple scene domains. The central problem in recovering intrinsic scene characteristics is that the information is confounded in the original light-intensity image: a single intensity value encodes all the characteristics of the corresponding scene point. Recovery depends on exploiting constraints, derived from assumptions about the nature of the scene and the physics of the imaging process.

I INTRODUCTION

Despite corsiderable progress in recent years, our understanding of the principles underlying visual perception remains primitive. Attempts to construct computer models for the interpretation of arbitrary scenes have resulted in such poor performance, limited range of abilities, and inflexibility that, were it not for the human existence proof, we might have been tempted long ago to conclude that high-performance, general-purpose vision is impossible. On the other hand, attempts to unravel the mystery of human vision, have resulted in a limited understanding of the elementary neurophysiology, and a wealth of phenomenological observations of the total system, but not, as yet, in a cohesive model of how the system functions. The time is right for those in both fields to take a broader view: those in computer vision might do well to look harder at the phenomenology of human vision for clues that might indicate fundamental inadequacies of current aproaches; these concerned with human vision might gain insights by thinking more about what information is sought, and how it might be obtained, from a computational point of view. This position has been strongly advocated for some time by Horn [18-20] and Marr [26-29] at MIT.

Current scene analysis systems often use pictorial features, such as regions of uniform intensity, or step charges in intensity, as an initial level of description and then jump directly to descriptions at the level of complete objects. The limitations of this approach are well known [4]: first, region-growing and edge-finding programs are unreliable in extracting the features that correspond to object surfaces because they have no basis for evaluating which intensity differences correspond to scene events significant at the level of objects (e.g., surface boundaries) and which do not (e.g., shadows). Second, matching pictorial features to a large number of object models is difficult and potentially combinatorially explosive because the feature descriptions are impoverished and lack invariance to viewing conditions. Finally, such systems cannot cope with objects for which they have no explicit model.

Some basic deficiencies in current approaches to machine vision are suggested when one examines the known behavior and competence of the human visual system. The literature abounds with examples of the ability of people to estimate characteristics intrinsic to the scene, such as color, orientation, distance, size, shape, or illumination, throughout a wide range of viewing conditions. Many experiments have been performed to determine the scope of so-called “shape constancy,” “size constancy,” and “color constancy” [13 and 14]. What is particularly remarkable is that consistent judgements can be made despite the fact that these characteristics interact strongly in determining intensities in the image. For example, reflectance can be estimated over an extraordinarily wide range of incident illumination: a black piece of paper in bright sunlight may reflect more light than a white piece in shadow, but they are still perceived as black and white respectively. Color also appears to remain constant throughout wide variation in the spectral composition of incident illumination. Variations in incident illumination are independently perceived: shadows are usually easily distinguished from changes in reflectance. Surface shape, too, is easily discerned regardless of illumination or surface markings: Yonas has experimentally determined that human accuracy in estimating local surface orientation is about eight degrees [37]. It is a worthwhile exercise at this point to pause and see how easily you can infer intrinsic characteristics, like color or surface orientation, in the world around you.

Copyright © 1978 by Academic Press, Inc.
All rights of reproduction in any form reserved.
ISBN 0-12-323550-2

The ability of humans to estimate intrinsic characteristics does not seem to require familiarity with the scene, or with objects contained therein. One can form descriptions of the surfaces in scenes unlike any previously seen, even when the presentation is as unnatural as a photograph. People can look at photomicrographs, abstract art, or satellite imagery, and make consistent judgements about relative distance, orientation, transparency, reflectance, and so forth. See, for example, Figure 1, from a thesis by Macleod [25].

Looking beyond the phenomenological aspects, one might ask what is the value of being able to estimate such intrinsic characteristics. Clearly, some information is valuable in its own right: for example, knowing the three-dimensional structure of the scene is fundamental to many activities, particularly to moving around and manipulating objects in the world. Since intrinsic characteristics give a more invariant and more distinguishing description of surfaces than raw light intensities, they greatly simplify many basic perceptual operations. Scenes can be partitioned into regions that correspond to smooth surfaces of uniform reflectance, and viewpoint-independent descriptions of the surfaces may then be formed [29]. Objects may be described and recognized in terms of collections of these elementary surfaces, with attributes that are characteristic of their composition or function, and relationships that convey structure, and not merely appearance. A chair, for example, can be described generically as a horizantal surface, at an appropriate height for sitting, and a vertical surface situated to provide back support. Previously unknown objects can be described in terms of invariant surface characteristics, and subsequently recognized from other viewpoints.

A concrete example of the usefulness of intrinsic scene information in computer vision can be obtained from experiments by Nitzan, Brain and Duda [30] with a laser rangefinder that directly measures distance and apparent reflectance. Figure 2a shows a test scene taken with a normal camera. Note the variation in intensity of the wall and chart due to variations in incident illumination, even though the light sources are extended and diffuse. The distance and reflectance for this scene is obtained by the rangefinder are shown in Figure 2b. The distance information is shown in a pictorial representation in which closer points appear brighter. Note that, except for a slight amount of crosstalk on the top of the cart, the distance image is insensitive to reflectance variations. The laser images are also entirely free from shadows.

Using the distance information, it is relatively straightforward to extract regions corresponding to flat or smooth surfaces, as in Fig. 2c, or edges corresponding to occlusion boundaries, as in Figure 2d, for example. Using reflectance information, conventional region- or edge-finding programs show considerable improvement in extracting uniformly painted surfaces. Even simple thresholding extracts acceptable surface approximations, as in Figure 2e.

Since we have three-dimensional information, matching is now facilitated. For example, given the intensity data of a planar surface that is not parallel to the image plane, we can eliminate the projective distortion in these data to obtain a normal view of this surface, Figure 2f. Recognition of the characters is thereby simplified. More generally, it is now possible to describe objects generically, as in the chair example above. Garvey [10] actually used generic descriptions at this level to locate objects in rangefinder images of office scenes.

The lesson to be learned from this example is that the use of intrinsic characteristics, rather than intensity values, alleviates many of the difficulties that plague current machine vision systems, and to which the human visual system is apparently largely immune.

The apparent ability of people to estimate intrinsic characteristics in unfamiliar scenes and the substantial advantages that such characteristics would provide strongly suggest that a visual system, whether for an animal or a machine, should be organized around an initial level of domain-independent processing, the purpose of which is the recovery of intrinsic scene characteristics from image intensities. The next step in pursuing this idea is to examine in detail the computational nature of the recovery process to determine whether such a design is really feasible.

In this paper, we will first establish the true nature of the recovery problem, and demonstrate that recovery is indeed possible, up to a point, in a simple world. We will then argue that the approach can be extended, in a straightforward way, to more realistic scene domains. Finally, we will discuss this paradigm and its implications in the context or current understanding of machine and human vision. For important related work see [29].

II THE NATURE OF THE PROBLEM

The first thing we must do is specify precisely the objectives of the recovery process in terms of input and desired output.

The input is one or more images representing light intensity values, for different viewpoints and spectral bands. The output we desire is a family of images for each viewpoint. In each family there is one image for each intrinsic characteristic, all in registration with the corresponding input images. We call these images “Intrinsic Images.” We want each intrinsic image to contain, in addition to the value of the characteristic at each point, explicit indications of boundaries due to discontinuities in value or gradient. The intrinsic images in which we are primarily interested are of surface reflectance, distance or surface orientation, and incident illumination. Other characteristics, such as transparency, specularity, luminosity, and so forth, might also be useful as intrinsic images, either in their own right or as intermediate results.

Figure 3 gives an example of one possible set of intrinsic images corresponding to a single, monochrome image of a simple scene. The intrinsic images are here represented as line drawings, but in fact would contain numerical values at every point. The solid lines show

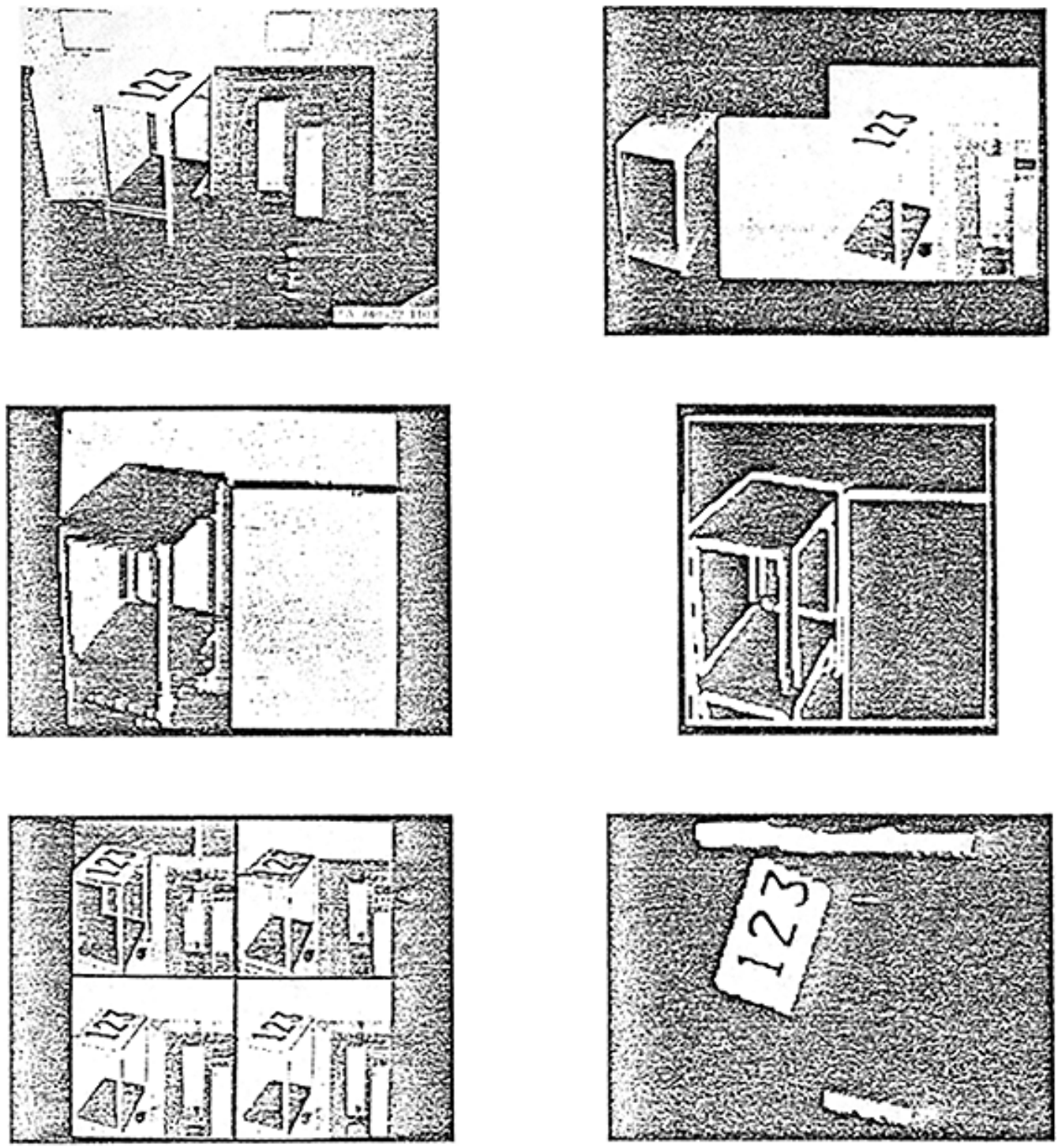
(a) CASTANOPSIS (X 3500)

(b) DRIMYS (X 3200)

(c) FLAX (X 1000)

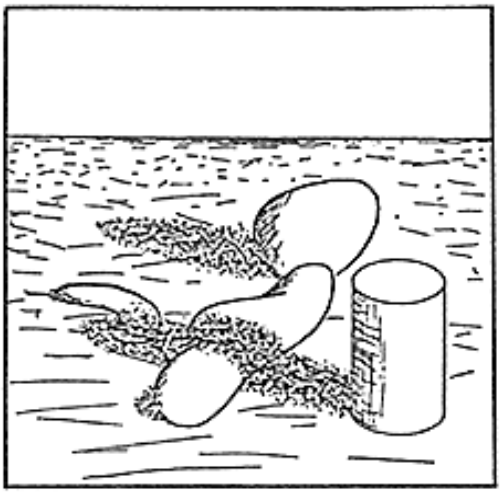
(d) WALLFLOWER (X 1800)

Figure 1 Photomicrographs of pollen grains (Macleod [20])



(a) THRESHOLDING REFLECTANCE (b) CORRECTED VIEW OF CART TOP

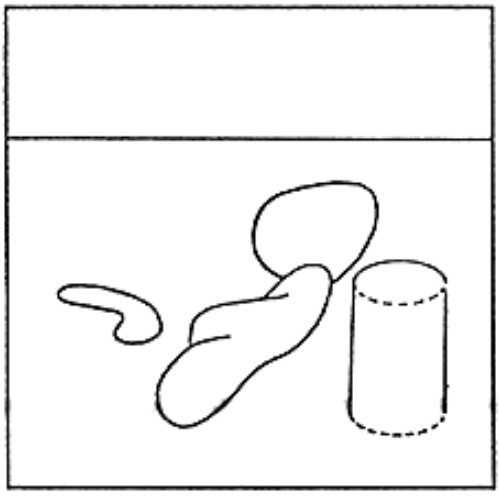
Figure 2 Experiments with a laser rangefinger



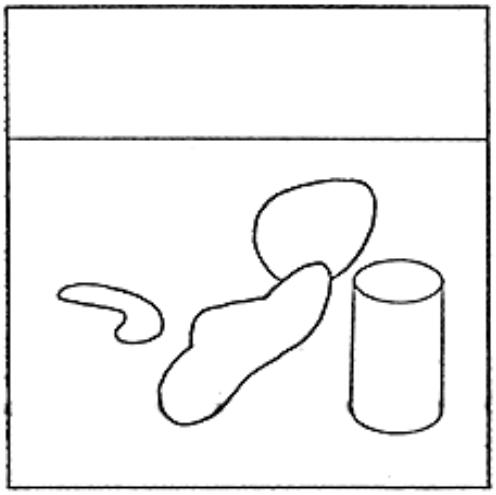
(a) ORIGINAL SCENE

Figure 3 A set of intrinsic images derived from a single monochrome intensity image

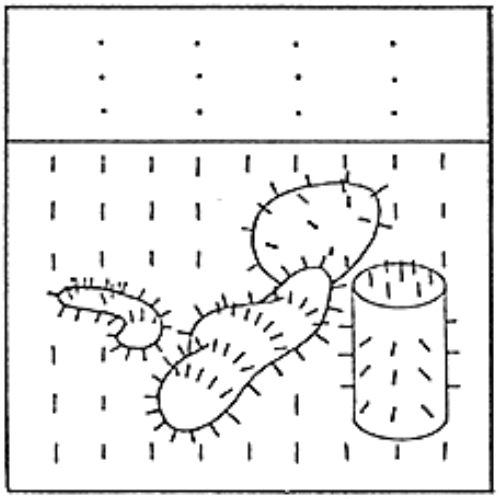
The images are depicted as line drawings, but, in fact, would contain values at every point. The solid lines in the intrinsic images represent discontinuities in the scene characteristic; the dashed lines represent discontinuities in its derivative.



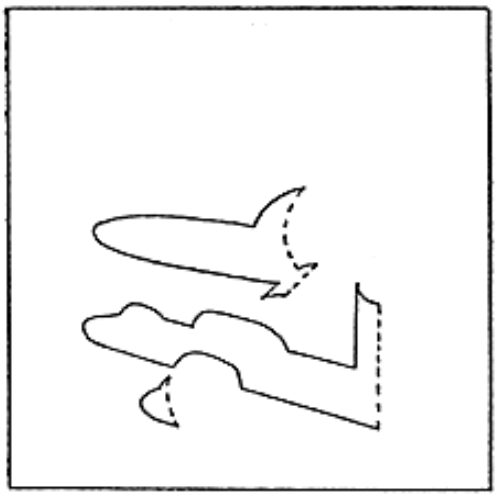
(b) DISTANCE



(c) REFLECTANCE



(d) ORIENTATION (VECTOR)



(e) ILLUMINATION

discontinuities in the represented characteristic, and the dashed lines show discontinuities in its gradient. In the input image, intensities correspond to the reflected flux received from the visible points in the scene. The distance image gives the range along the line of sight from the center of projection to each visible point in the scene. The orientation image gives a vector representing the direction of the surface normal at each point. It is essentially the gradient of the distance image. The short lines in this image are intended to convey to the reader the surface orientation at a few sample points. (The distance and orientation images correspond to Marr's notion of a 2.5D sketch [29].) It is convenient to represent both distance and orientation explicitly, despite the redundancy, since some visual cues provide evidence concerning distance and other evidence concerning orientation. Moreover, each form of information may be required by some higher-level process in interpretation or action. The reflectance image gives the albedo (the ratio-of total reflected to total incident illumination) at each point. Albedo completely describes the reflectance characteristics for lambertian (perfectly diffusing) surfaces, in a particular spectral band. Many surfaces are approximately lambertian over a range of viewing conditions. For other types of surface, reflectance depends on relative directions of incident rays, surface normal and reflected rays. The illumination image gives the total light flux incident at each point. In general, to completely describe the incident light it is necessary to give the incident flux as a function of direction. For point light sources, one image per source is sufficient, if we ignore secondary illumination by light scattered from nearby surfaces.

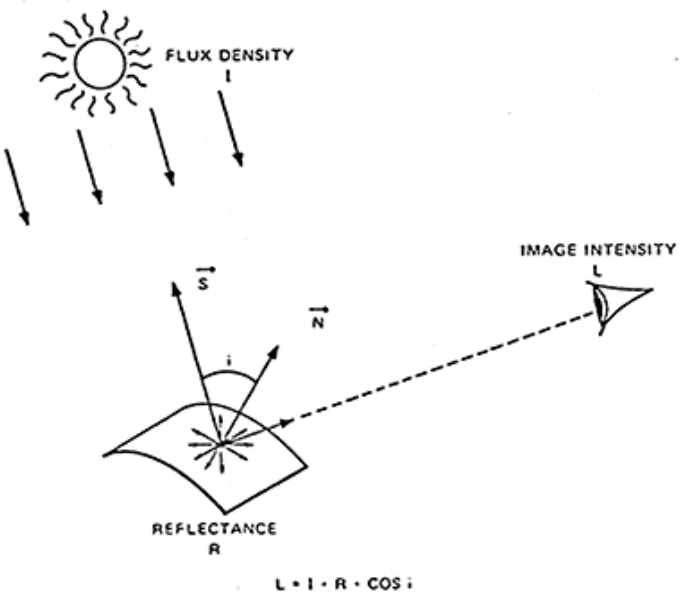


Figure 4 An ideally diffusing surface

When an image is formed, by a camera or by an eye, the light intensity at a point in the image is determined mainly by three factors at the corresponding point in the scene: the incident illumination, the local surface reflectance, and the local surface orientation. In the simple case of an ideally diffusing surface illuminated by a point source, as in Figure 4, for example, the image light intensity, L, is given by

$$L = I * R * \cos i \qquad (1)$$

where I is intensity of incident illumination, R is reflectivity of the surface, and i is the angle of incidence of the illumination [20].

The central problem in recovering intrinsic scene characteristics is that information is confounded in the light-intensity image: a single intensity value encodes all the intrinsic attributes of the corresponding scene point. While the encoding is deterministic and founded upon the physics of imaging, it is not unique: the measured light intensity at a single point could result from any of an infinitude of combinations of illumination, reflectance, and orientation.

We know that information in the intrinsic images completely determines the input image. The crucial question is whether the information in the input image is sufficient to recover the intrinsic images.

III THE NATURE OF THE SOLUTION

The only hope of decoding the confounded information is, apparently, to make assumptions about the world and to exploit the constraints they imply. In images of three-dimensional scenes, the intensity values are not independent but are constrained by various physical phenomena. Surfaces are continuous in space, and often have approximately uniform reflectance. Thus, distance and orientation are continuous, and reflectance is constant everywhere in the image, except at edges corresponding to surface boundaries. Incident illumination, also, usually varies smoothly. Step changes in intensity usually occur at shadow boundaries, or surface boundaries. Intrinsic surface characteristies are continuous through shadows. In man-made environments, straight edges frequency correspond to boundaries of planar surfaces, and ellipses to circles viewed obliquely. Many clues of this sort are well known to psychologists and artists. There are also higher-level constraints based on knowledge of specific objects, or classes of object, but we shall not concern ourselves with them here, since our aim is to determine how well images can be interpreted without object-level knowledge.

We contend that the constraints provided by such phenomena, in conjunction with the physics of imaging, should allow recovery of the intrinsic images from the input image. As an example, look carefully at a nearby painted wall. Observe that its intensity is not uniform, but varies smoothly. The variation could be due, in principle, to variations in reflectance, illumination, orientation, or any combination of them. Assumptions of continuity immediately rule out the situation of a smooth intensity variation arising from cancelling random variations in illumination, reflectance, and orientation since surfaces are assumed to

be uniform in reflectance, the intensity variation must thus be due to a smooth variation in illumination or surface shape. The straight edge of the wall suggests, however, that the wall is planar, and that the variation is in illumination only. To appreciate the value of this constraint, view a small central portion of the wall through a tube. With no evidence from the edge, it is difficult to distinguish whether the observed shading is due to an illumination gradient on a planar surface, or to a smooth surface curving away from the light source.

The tube experiment shows that while isolated fragments of an image have inherent ambiguity, interactions among fragments resulting from assumed constraints can lead to a unique interpretation of the whole image. Of course, it is possible to construct (or occasionally to encounter) scenes in which the obvious assumptions are incorrect -- for example, an Ames roon (see [13] for an illustration). In such cases, the image will be misinterpreted, resulting in an illusion. The Ames illusion is particularly interesting because it shows the lower-level interpretation, of distance and orientation, dominating the higher-level knowledge regarding relative sizes of familiar objects, and even dominating size constancy. Fortunately, in natural scenes, as commonly encountered, the evidence is usually overwhelmingly in favor of the correct interpretation.

We have now given the flavor of the solution, but with many of the details lacking. Our current research is aimed at making the underlying ideas sufficiently precise to implement a computational model. While we are far from ready to attack the full complexity of the real world, we can give a fairly precise description of such a model for recovering intrinsic characteristics in a limited world. Moreover, we can argue that this model may be extended incrementally to handle more realistic scenes.

IV SOLUTION FOR A SIMPLE WORLD

A. Methodology

To approach the problem systematically, we select an idealized domain in which a simplified physics holds exactly, and in which explicit contraints on the nature of surfaces and illuminants. From these assumptions, it is possible to enumerate various types of scene fragments and determine the appearance of their corresponding image fragments. A catalog of fragment appearances and alternative interpretations can thus be compiled (in the style of Huffman [21] and Waltz [34]).

We proceed by constructing specific scenes that satisfy the domain assumptions, synthesizing corresponding images of them, and then attempting to recover intrinsic characteristics from the images, using the catalog and the domain knowledge. (By displaying synthetic images, we could check that people can interpret them adequately. If they cannot, we can discover oversimplifications by comparing the synthetic images to real images of similar scenes.)

B. Selection of a Domain

Specifications for an experimental domain must include explicit assumptions regarding the scene, the illumination, the viewpoint, the sensor, and the image-encoding process. The initial domain should be sufficiently simple to allow exhaustive enumeration of its constraints, and complete cataloging of appearances. It must, however, be sufficiently complex so that the recovery process is non-trivial and generalizable. A domain satisfying these requirements is defined as follows:

* Objects are relatively smooth, having surfaces over which distance and orientation are continuous. That is, there are no sharp edges or creases.
* Surfaces are lambertian reflectors, with constant albedo over them. That is, there are no surface markings and no visible texture.
* Illumination is from a distant point source, of known magnitude and direction, plus uniformly diffuse background light of known magnitude (an approximation to sun, sky, and scattered light). Local secondary illumination (light reflected from nearby objects) is assumed to be negligible. (See Figure 5.) ;
* The image is formed by central projection onto a planar surface. Only a single view is available (no stereo or motion parallax). The scene is viewed from a general position (incremental changes in viewpoint do not change the topology of the image) .
* The sensor measures reflected flux density. Spatial and intensity resolution are sufficiently high that quartization effects may be ignored. Sensor noise is also negligible.

Such a domain might be viewed as an approximation of a world of colored Play-Doh objects in which surfaces are smooth, reflectance is uniform for each object, there is outdoor illumination, and the scene is imaged by a tv camera. The grossest approximations, perhaps, are the assumptions about illumination, but they are substantially more realistic than the usual single-point-source model, which renders all shadowed regions perfectly black.

For this domain, our objective is to recover intrinsic images of distance, orientation, reflectance, and illumination.

C. Describing the Image

Elementary physical considerations show that a portion of a surface that is continuous in visibility, distance, orientation, and incident illumination, and has uniform reflectance, maps to a connected region of continuous intensity in the image. Images thus consist of regions of smoothly varying intensity, bounded by step discontinuities. In our domain, reflectance is constant over each surface, and there are two states of illumination, corresponding to sun and shadow. Image regions therefore correspond to areas of surface with a particular state of illumination, and the boundaries corresponding

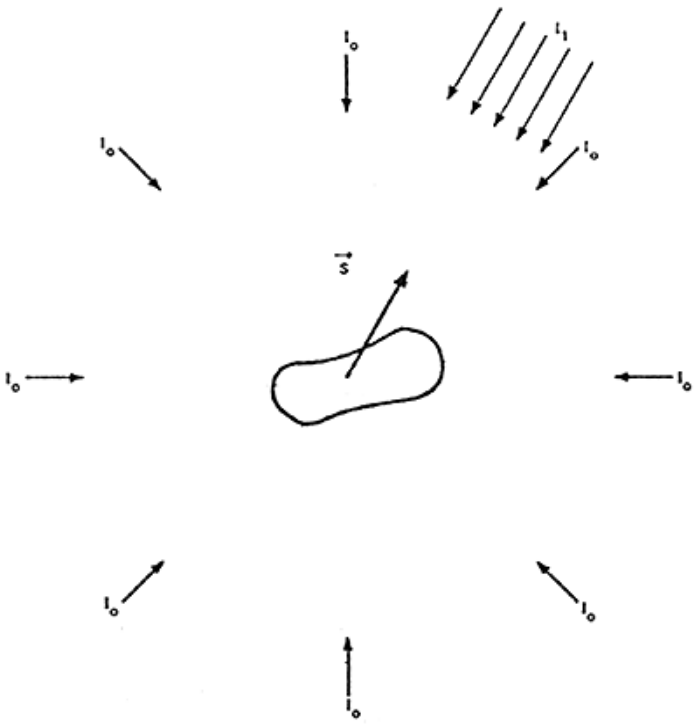


Figure 5 Sun and sky illumination model

to occluding (extremal) boundaries of surfaces, or to the edges of shadows. There are also junctions where boundaries meet. Figure 6b shows the regions and edges for the simple scene of Figure 3.

To be quantitative, we assume image intensity is calibrated to give reflected flux density at the corresponding scene point. Reflected flux density is the product of integrated incident illumination, I, and reflectance (albedo), R, at a surface element. Thus,

$$L = I * R \qquad (2)$$

The reflected light is distributed uniformly over a hemisphere for a lambertian surface. Hence, image intensity is independent of viewing direction. It is also independent of viewing distance, because although the flux density received from a unit area of surface decreases as the inverse square of distance, the surface area corresponding to a unit area in the image increases as the square of distance.

In shadowed areas of our domain, where surface elements are illuminated by uniform diffuse illumination of total incident flux density I0, the image intensity is given by

$$L = IO * R \qquad (3)$$

When a surface element is illuminated by a point source, such that the flux density is I1, from a direction specified by the unit vector, S, the incident flux density at the surface is I1 * N.S, where N is the unit normal to the surface, and . is the vector dot product. Thus,

$$L = I1 * N.S * R \qquad (4)$$

In directly illuminated areas of the scene, image intensity, L, is given by the sum of the diffuse and point-source components:

$$L = (I0 + I1 + N.S) * R \qquad (5)$$

From the preceding sections, we are not in a position to describe the appearance of image fragments in our domain, and then to derive a catalog.

1. Regions

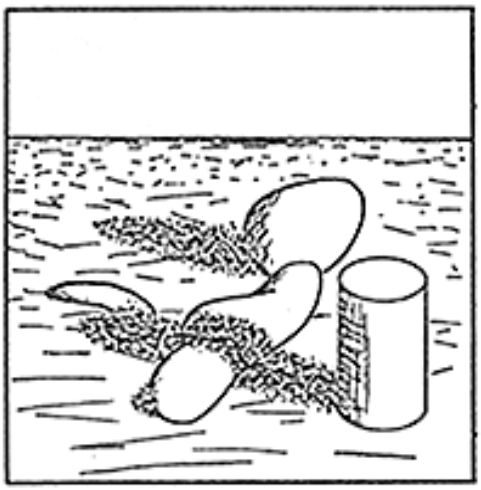
For a region corresponding to a directly illuminated portion of a surface, since R, I0, and I1 are constant, any variation in image intensity is due solely to variation in surface orientation. For a region corresponding to a shadowed area of surface, intensity is simply proportional to reflectance, and hence is constant over the surface.

We now catalog regions by their appearance. Regions can be classified initially according to whether their intensities are smoothly varying, or constant. In the former case, the region must correspond to a nonshadowed, curved surface with constant reflectance and continuous depth and orientation. In the latter case, it must correspond to a shadowed surface. (An illuminated planar surface also has constant intensity, but such surfaces are excluded from our domain.) The shadowing may be due either to a shadow cast upon it, or to its facing away from the point source. The shape of a shadowed region is indeterminable from photometric evidence. The surface may contain bumps or dents and may even contain discontinuities in orientation and depth across a self-occlusion, with no corresponding intensity variations in the image.

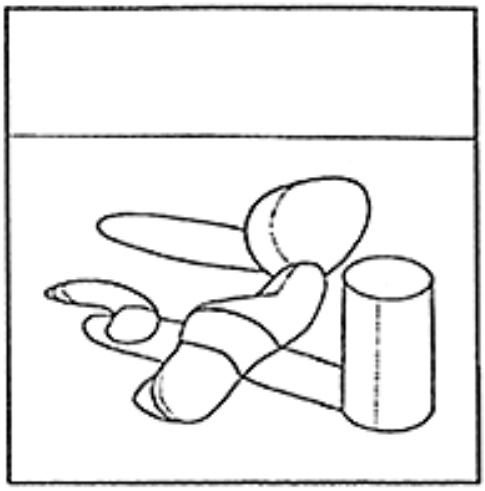
2. Edges

In the same fashion as for regions, we can describe and catalog region boundaries (edges). An edge should not be considered merely as a step change in image intensity, but rather as an indication of one of several distinct scene events. In our simple world, edges correspond to either the extremal boundary of a surface (the solid lines in Figure 3b), or to the boundary of a cast shadow (the solid lines in Figure 3e). The "terminator" line on a surface, where there is a smooth transition from full illumination to self-shadowing (the dashed lines in Figure 3e), does not produce a step change in intensity

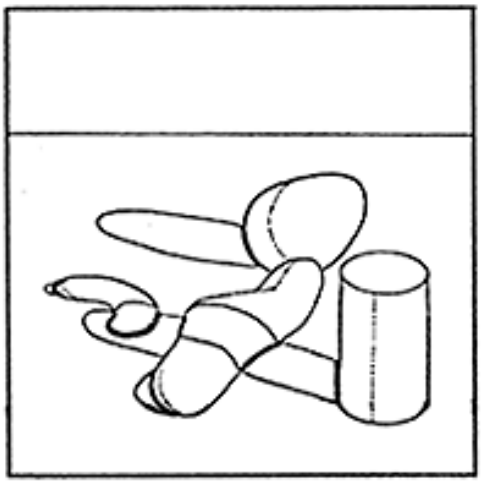
The boundary of a shadow cast on a surface indicates only a difference in incident illumination: the intrinsic characteristics of the surface are continuous across it. As we observed earlier, the shadowed region is constant in intensity, and the illuminated region has an intensity gradient that is a function of the surface orientation. The shadowed region is necessarily darker than the illuminated one.



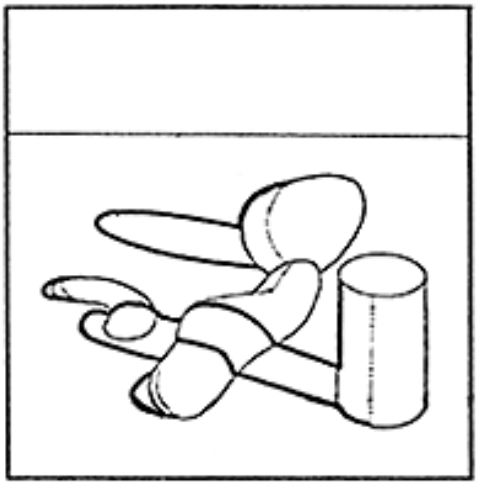
(a) ORIGINAL SCENE



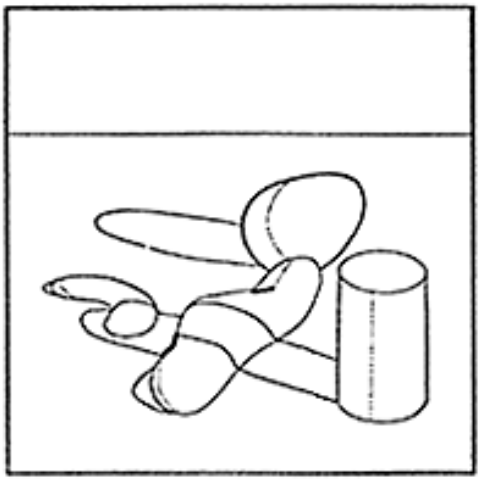
(b) INPUT INTENSITY IMAGE



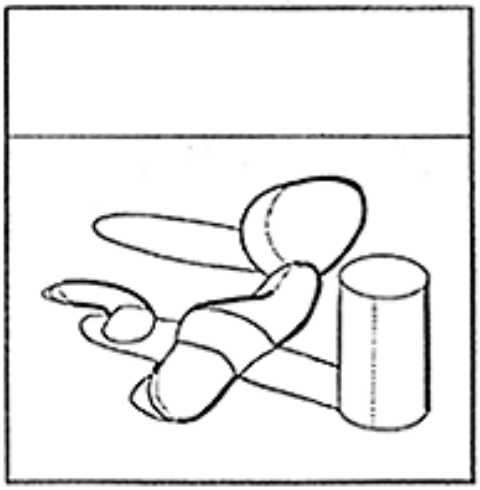
(c) LA: CONSTANT, LB: CONSTANT



(d) LA: CONSTANT, LB: VARYING



(e) LA: CONSTANT, LB: TANGENT



(f) LA: VARYING, LB: TANGENT

Figure 6 Initial classification of edges in an example scene.

An extremal boundary is a local extremum of the surface from the observer's point of view, where the surface turns away from him. Here one surface occludes another, and all intrinsic characteristics may be discontinuous. In our world, it is assumed that depth and orientation are always discontinuous. Reflectance is constant on each side of the edge, and will only be continuous across it if the two surfaces concerned have identical reflectance, or when a single surface occludes itself.

A very important condition results from the fact that an extremal boundary indicates where a smooth surface turns away from the viewer: the orientation is normal to the line of sight, and to the tangent to the boundary in the image. Hence the absolute orientation of the occluding surface can be determined along an extremal boundary. The boundary must first be identified as extremal, however.

The regions on either side of an extremal boundary are independently illuminated. When both are in shadow, they both have constant intensity, and the ratio of intensities is equal to the ratio of the reflectances: it is not possible to determine from local intensity information which region corresponds to the occluding surface.

When a region is illuminated, intensity varies continuously along its side of the boundary. We noted that the image of an extremal boundary tells us precisely the orientation of the occluding surface at points along it. The orientation, together with tbe illumination flux densities, I0 and I1, and the image intensity, L, can be used in Equation (5) to determine reflectance at any point on the boundary. For a true extremal boundary, our assumption of uniform reflectance means that estimates of reflectance at all points along it must agree. This provides a basis for recognizing an occluding surface by testing whether reflectances derived at several boundary points are consistent. We call this test the tangency test, because it depends upon the surface being tangential to the line of sight at the boundary. This test is derived by differentiating the logarithm of Equation (5):

$$dL/L = dR/R + (I1*dN.s)/(I0 + I1*N.s) \quad (6)$$

The vector dN is the derivative of surface orientation along the edge, and may be determined from the derivative of edge direction. The derivatives dL and dR are taken along the edge. Equation (6) may be rewritten to give dR/R explicitly in terms of L, dL, N, dN and the constants I0, I1, and S. The tangency condition is met when dR/R is zero. The tangency condition is a powerful constraint that can be exploited in further ways, which we will discuss later.

Strictly speaking, where we have referred to derivatives here, we should have said "the knit of the derivative as the edge is approached from the side of the region being tested." Clearly, the tests are not applicable at gaps, and the tangency test is not applicable where the edge is discontinuous in direction.

We can now catalog edges by their appearances, as we did for regions. Edges are classified according to the appearance of the regions on either side in the vicinity of the edge. This is done by testing intensity values on each side for constancy, as before, and for satisfaction of the tangency test, and by testing relative intensities across the edge. Table 1 catalogs the possible appearances and interpretations of an edge between two regions, A and B.

In this table, "Constant" means constant intensity along the edge, "Tangency" means that the tangency condition is met, and

Table 1 The Nature of Edges

Region Intensities LA	Region Intensities LB	Edge Type	Region Types	Intrinsic Edges Intrinsic Values D	N	R	I
Constant	Constant	Occluding sense unknown	A B shadowed	EDGE	EDGE	EDGE RA RB	IA IB
Constant	Varying	1 Shadow	A shadowed B illuminated		NB.S	RA RB	EDGE IA IB
		2 A occludes B	A shadowed B illuminated	EDGE DA DB	EDGE NA	EDGE RA	EDGE IA
Varying	Varying	Inconsistent with domain					
Constant	Tangency	B occludes A	A shadowed B illuminated	EDGE DA DB	EDGE NB	EDGE RA RB	EDGE IA IB
Varying	Tangency	B occludes A	A B illuminated	EDGE DA DB	EDGE NB	EDGE RB	EDGE IB IA
Tangency	Tangency	Not seen from general position					

"Varying" means that neither of these tests succeeds. The entry "EDGE" denotes a discontinuity in the corresponding intrinsic attribute, with the same location and direction as the corresponding image intensity edge. The magnitude and sense of the discontinuity are unknown, unless otherwise shown. Where the value of an intrinsic attribute can be determined from the image (see Section IV.D), it is indicated by a term of the form RA, RB, DA, etc. (These terms denote values for reflectance, R; orientation, vector N; distance, D; and incident flux density, I; for the regions A and B, in the obvious way.) Where only a constraint on values is known, it is indicated by an inequality. There is a special situation, concerning case 1 of the second type of edge, in which a value can be determined for NB.S, but not for NB itself.

Note that, from the types of intensity variations, edges can be interpreted unambiguously, except for two cases: namely, the sense of the occluding edge between two shadowed regions, and the interpretation of an edge between illuminated and shadowed regions when the tangency test fails. Figure 6 illustrates the classification of edges according to tbe catalog for our example scene.

3. Junctions

Since the objects in our domain are smooth, there are no distinguished points on surfaces. Junctions in the image, therefore, are viewpoint dependent. There are just two classes of junction, both resulting from an extremal boundary, and both appearing as a T-shape in the image (see Figure 7).

The first type of junction arises when one object partially occludes a more distant boundary, which may be either a shadow edge or an extremal edge. The crossbar of the junction is thus an extremal edge of an object that is either illuminated or shadowed. The boundary forming the stem lies on the occluded object and its edge type is unconstrained.

The second type of junction arises when a shadow cast on a surface continues around behind an extremal boundary. In this case, the crossbar is again an extremal edge, half in shadow, while the stem is a shadow edge lying on the occluding object.

Note that in both cases the crossbar of the T corresponds to a continuous extremal boundary. Hence the two edges forming the crossbar are continuous, occluding, and have the same sense.

The T junctions provide constraints that can sometimes resolve the ambiguities in the edge table above. Consider the cases as follows: if all the regions surrounding the T are shadowed, the edge table tells us that all the edges are occluding, but their senses are ambiguous. The region above the crossbar, however, must be occluding the other two, otherwise the continuity of the crossbar could be due to an accident of viewpoint. If one or more of the regions is illuminated, the occlusion sense of the crossbar is immediately determined from the tangency test. Thus we can always determine the nature of the two edges forming the crossbar of the T, even wnen they may have been ambiguous according to the edge table.

If the region above the crosbar is the occluder, we have the first type of T junction, and can say no more about the stem than the edge tests give us. Otherwise, we have the second type (a shadow edge cast over an extremal boundary), and any ambiguity of the stem is now resolved.

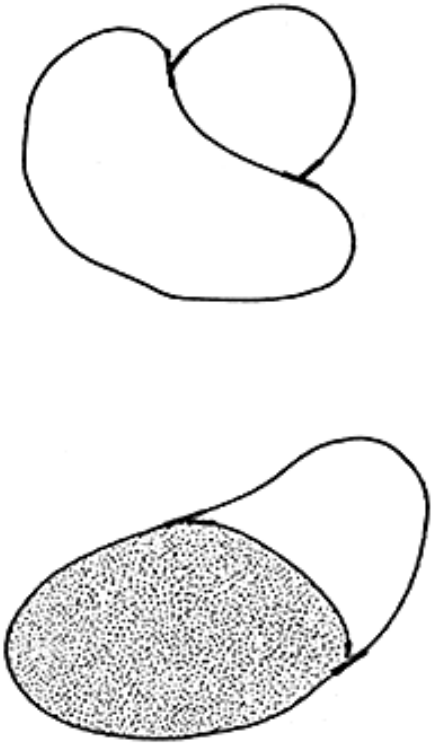


Figure 7 Two types of T-junction

D. Recovery Using the Catalog

The ideal goal is to recover all intrinsic scene characteristics, exactly, everywhere in an image that is consistent with our domain. In this section, we outline the principles of recovery using the catalog and address the issue of how nearly our goal can be attained. The following section will describe a detailed computational model of the recovery process.

The recovery process has four main steps:

(1) Find the step edges in the input intensity image.

(2) Interpret the intrinsic nature of the regions and edge elements, according to the catalog. Interpretation is based on the results of constancy and tangency tests.

(3) Assign initial values for intrinsic characteristics along the edges, based on the interpretations.

(4) Propagate these "boundary" values into the interiors of regions, using

continuity assumptions. This step is analogous to the use of relaxation methods in physics for determining distributions of temperature or potential over a region based on boundary values.

For pedagogical purposes, in this section only, we assune that it is possible to extract a perfect line drawing from the intensity image. The recovery process described later does not depend on this assumption, because: it perfects the line drawing as an integral part of its interpretation. Let us now consider the ultimate limitations of the recovery paradigm in our simple domain.

Shadowed and directly illuminated areas of the image are distinguished immediately, using the constancy test. Reflectance everywhere in shadowed areas is then given by Equation (3).

The orientation of a region corresponding to an illuninated surface can be determined along extremal boundaries identified by the tangency test. Reflectance of this region can then be determined at the boundary by Equation (5), and thus throughout the region based on the assumption that reflectance is constant over a surface.

So far, recovery has been exact; the intrinsic values and edges that can be exactly inferred from intensity edges are shown in Table 1. Surface orientation within illuminated regions bounded, at least in part, by extremal edges can be reasonably estimated, as follows: Equation (5) can be solved, knowing L, R, I0, and I1, for N.S, the cosine of the angle of incidence of the direct component of illumination, at each point. This does not uniquely determine the orientation, which has two degrees of freedom, but it does allow reasonable estimates to be obtained using the assumption of smoothness of surfaces and the known orientation at extremal boundaries to constrain the other degree of freedom. Two examples of how this reconstruction can be done are given by the work of Horn [19] and Woodham [35] on "shape fron shading."

Orientation can be integrated to obtain relative distance within the regions, and the tangency test gives distance ordering across the boundary.

Since a shadowed region of surface appears uniform in the intensity image, its shape cannot be determined from shading information. A plausible guess can be made, however, by interpolating in from points of known orientation, using the smoothness assumption. This can be done only if at least part of the boundary of the shadowed region can be interpreted as extremel boundary (e.g., using T-junctions), or as a shadow edge with the shape on the illuminated side known.

Not surprisingly, little can be said about regions, shadowed or illuminated, with no visible portions of boundary identifiable as extremal (e.g. a region seen through a hole, or shadowed, with no T-junctions). It is still reasonable to attempt to interpret such inherently ambiguous situations, but it is then necessary to introduce further, and perhaps less general, assumptions. For example: an object is predominantly convex, so the sense of an occlusion can be guessed locally from the shape of the boundary; the brightest point on an illuminated surface is probably oriented with its normal pointing at the light source, providing a boundary condition for determining the surface reflectance and its shape from shading. Of course, such assumptions must be subordinate to harder evidence, when it is available.

We conclude that, in this limited domain, unambiguous recovery of intrinsic characteristics at every point in an image is not generally possible, primarily because of the lack of information in some regions of the intensity image. Thus, in some cases, we must be content with plausible estimates derived from assumptions about likely scene characteristics. When these assumptions are incorrect, the estimates will be wrong, in the sense that they will not correspond exactly to the scene; they will, however, provide an interpretation that is consistent with the evidence available in the intensity image, and most of the time this interpretation will be largely correct.

Though perfect recovery is unattainable, it is remarkable how much can be done considering the weakness (and hence generality) of the assumptions, and the limited number of cues available in this domain. We used no shape prototypes, nor object models, and made no use of any primary depth cues, such as stereopsis, motion parallax, or texture gradient. Any of these sources, if available, could be incorporated to improve performance by providing information where previously it could only be guessed (for example, texture gradient could eliminate shape ambiguity in shadows).

V A COMPUTATIONAL MODEL

We now propose a detailed computational model of the recovery process. The model operates directly on the data in a set of intrinsic images and uses parallel local operations that codify values in the images to make them consistent with the input image and constraints representing the physical assumptions about imaging and the world.

A. Establishing Constraints

Recovery begins with the detection of edges in the intensity image. If quantization and noise are assumed negligible in the domain, we can easily distinguish all step discontinuities, and hence generate a binary image or intensity edges, each with an associated direction. This image will resemble a perfect line drawing, but despite the ideal conditions, there can still be gaps where intensities on two sides or a boundary happen to be identical. Usually this will occur only at a single isolated point -- for example, at a point where the varying intensities on the two sides of an occlusion boundary simultaneously pass through the same value. Complete sections of a boundary may also occasionally be invisible -- for example, when a shadowed body occludes itself. Our recovery process is intended to cope with these imperfections, as will be seen later.

Given the edge image, the next step is to interpret the intrinsic nature of the edge elements according to the edge table. Interpretation is based on the results of two tests, constancy and tangency applied to the intensities of the regions immediately adjacent

to an edge element. The constancy test is applied by simply checking whether the gradient of intensity is zero. The tangency test is applied in its differential form by checking whether the derivative of estimated reflectance, taken along the edge, is zero.

The resulting edge interpretations are used to initialize values and edges in the intrinsic images in accordance with the edge table. The table specifies, for each type of intensity edge, the intrinsic images in which corresponding edge elements should be inserted. Values are assigned to intrinsic image points adjacent to edges, as indicated in the table. For example, if an intensity edge is interpreted as an occlusion, we can generate an edge in the distance and orientation images, and initialize orientation and reflectance images at points on the occluding side of the boundary.

When the edge interpretation is ambiguous, we make conservative initializations, and wait for subsequent processing to resolve the issue. In the case of an extremal boundary separating two shadowed regions, this means assigning a discontinuity in distance, orientation and reflectance, but not assuming anything else about orientation or relative distance. In the case of ambiguity between a shadow edge and a shadowed occluding surface, we assume discontinuities in all characteristics, and that the illumination and reflectance of the shadowed region are known. It is better to assume the possible existence of an edge, when unsure, because it merely decouples the two regions locally; they may still be related through a chain of constraints along some other route.

For points at which intrinsic values have not been uniquely specified, initial values are assigned that are reasonable on statistical and psychological grounds. In the orientation image, the initial orientation is assigned as N0, the orientation pointing directly at the viewer. In the illumination image, areas of shadow, indicated by the constancy test, are assigned value I0. The remaining directly illuminated points are assigned value I1*N0.S In shadowed areas, reflectance values are assigned as L/I0, and in illuminated areas, they are assigned L/(I0 + I1*N0.S). Distance values are more arbitrary, and we assign a uniform distance, DO everywhere. The choice of default values is not critical, they simply serve as estimates to be improved by the constraint satisfaction processes.

Following initialization, the next step is to establish consistency of intrinsic values and edges in the images. Consistency within an individual image is governed by simple continuity and limit constraints. In the reflectance image, the constraint is that reflectance is constant -- that is, its gradient must be defined and be zero everywhere, except at a reflectance edge. Reflectance is additionally constrained to take values between 0 and 1. Orientation values are also constrained to be continuous, except at occlusion edges. The vectors must be unit vectors, with a positive component in the direction of the viewpoint. Illumination is positive and continuous, except across shadow boundaries. In shadowed regions, it must be ccnstant, and equal to I0. Distance values must be continuous everywhere -- that is, their gradient must be defined and finite, except across occlusion edges. Where the sense of the occlusion is known, the sense of the discontinuity is constrained appropriately. Distance values must always be positive.

All these constraints involve local neighborhoods, and can thus be implemented via asynchronous parallel processes. The continuity constraints, in particular, might be implemented by processes that simply ensure that the value of a characteristic at a point is the average of neighboring values. Such processes are essentially solving Laplace's equation by relaxation.

The value at a point in an intrinsic image is related not only to neighboring values in the same image, but also to values at the corresponding point in the other images. The primary constraint of this sort is that image intensity is everywhere the product of illumination and reflectance, as in Equation (2). Incident illumination is itself a sum of terms, one for each source. This may conveniently be represented by introducing secondary intrinsic images for the individual sources. The image representing diffuse illumination is constant, with value I0, while that for the point source is I1*N.S, where N.S is positive and the surface receives direct illumination, and zero elsewhere. These constraints tie together values at corresponding points in the input intensity, reflectance, primary and secondary illumination, and orientation images. The orientation and distance images are constrained together by the operation of differentiation.

B. Achieving Consistency

Consistency is attained when the intra- and inter-image constraints are simultaneously satisfied. That occurs when values and gradients everywhere are consistent with continuity assumptions and constraints appropriate to the type of image, and the presence or absence of edges. The intraimage constraints ensure that intrinsic characteristics vary smoothly, in the appropriate ways. Such constraints are implicit in the domain description, and are not usually made so explicit in machine vision systems. The inter-image constraints ensure that the characteristics are also consistent with the physics of imaging and the input intensity image. It is these constraints that permit an estimate of surface shape from the shading within regions of smoothly varying intensity [19].

Consistency is achieved by allowing the local constraint processes, operating in parallel, to adjust intrinsic values. The explicitly determined values listed in the initialization table, however, are treated as boundary conditions and are not allowed to change. As a result, information propagates into the interior of regions from the known boundaries.

The initial assignment of intrinsic edges is, as we have already noted, imperfect: edges found in the intensity image may contain gaps, introducing corresponding gaps in the intrinsic edges; certain intrinsic edges are not visible in the intensity image -- for example, self-occlusion in shadow; some intrinsic cases were assumed in the interests of conservatism, but they may be incorrect. From the recovered intrinsic values, it may be clear where futher

edges should be inserted in (or deleted from) the corresponding intrinsic image. For example, when the gradient in an image becomes very high, insertion of an edge element is indicated, and, conversely, when the difference in values across an edge becomes very small, deletion is indicated. Insertion and deletion must operate within existing constraints. In particular, edge elements cannot be manipulated within a single image in isolation, because a legal interpretation must always be maintained. For example, in our world, occlusion implies edges in distance and orientation simultaneously. Within an intrinsic image, continuity of surfaces implies continuity of boundaries. Hence, the decision to insert must take neighboring edge elements into consideration.

Constraints and boundary conditions are dependent upon the presence or absence of edges. For example, if a new extremal edge is inserted, continuity of distance is no longer required at that point, and the orientation on one side is absolutely determined. Consequently, when edge elements are inserted or deleted, the constraint satisfaction problem is altered. The constraint and edge modification processes run continuously, interacting to perfect the original interpretation and recover the intrinsic characteristics. Figures 8 and 9 summarize the overall organization of images and constraints.

So far we have not mentioned the role of junctions in the recovery process. At this point, it is unclear whether junctions need to be treated explicitly since the edge constraints near a confluence of edges will restrict relative values and interpretations. Junctions could also be used in an explicit fashion. When a T-configuration is detected, either during initialization or subsequently, the component edges could be interpreted via a junction catalog, which would provide more specific interpretations than the edge table. Detection of junctions could be performed in parallel by local processes, consistent with the general organization of the system. The combinatorics of junction detection are much worse than those of edge detection, and have the consequence that reliability of detection is also worse. For these reasons, it is to be hoped that explicit reliance upon junctions will not be necessary.

The general structure of the system is clear, but a number of details remain to be worked out. These include: how to correctly represent and integrate inter- and intra-image constraints; how to insert and delete edge points; how to correctly exploit junction constraints; how to ensure rapid convergence and stability. Although we do not yet have a computer implementation of the complete model, we have been encouraged by experiments with some of the key components.

We have implemented a simple scheme that uses a smoothness constraint to estimate surface orientation in region interiors from boundary values. The constraint is applied by local parallel processes that replace each orientation vector by the average of its neighbors. The surface reconstructed is a quadratic function of the image coordinates. It is smooth, but not uniformly curved, with its boundary lying in a plane. It appears possible to derive more complex continuity constraints that result in more uniformly curved surfaces, interpolating, for example, spherical and cylindrical surfaces from their silhouettes.

The above smoothing process was augmented with another process that simultaneously adjusts the component of orientation in the direction of steepest intensity gradient to be consistent with observed intensity. The result is a cooperative "shape from shading" algorithm, somewhat different from Woodham's [35]. The combined algorithm has the potential for reconstructing plausible surface shape in both shadowed and directly illuminated regions.

VI EXTENDING THE THEORY

Our initial domain was deliberately oversimplified, partly for pedagogic purposes and partly to permit a reasonably exhaustive analysis. The approach of thoroughly describing each type of scene event and its appearance in the image, and then inverting to form a catalog of interpretations for each type of image event, is more generally applicable. Results so far appear to indicate that while ambiguities increase with increasing domain complexity, available constraints also increase proportionately. Information needed to effect recovery of scene characteristics seems to be available; it is mainly a matter of thinking and looking hard enough for it.

In this section, we will briefly describe some of the ways in which the restrictive assumptions of our initial domain can be relaxed, and the recovery process correspondingly augmented, to bring us closer to the real world.

The assumption of coutinuous, noise-free encoding is important for avoiding preoccupation with details of implementation, but it is essential for a realistic theory to avoid reliance upon it. With these assumptions, problems of edge detection are minimized, but, as we noted earlier, perfect line drawings are not produced. Line drawings conventionally correspond to surface outlines, which may not be visible everywhere in the image. The recovery process we described, therefore, incorporated machinery for inserting and deleting edges to achieve surface integrity. Relaxing the assumption of ideal encoding will result in failure to detect some weak intensity edges and possibly the introduction of spurious edge points in areas of high intensity gradient. Insofar as these degradations are moderate, the edge-refinement process should ensure that the solution is not significantly affected.

The assumption of constant reflectance on a surface can be relaxed by introducing a new edge type -- the reflectance edge -- where orientation, distance, and illumination are still continuous, but reflectance undergoes a step discontinuity. Reflectance edges bound markings on a surface, such as painted letters or stripes. The features distinguishing the appearance of an illuminated reflectance edge are that the ratio of edge intensities across the edge is constant along it and equal to the ratio of the reflectances and the magnitude of intensity gradient across the edge is also equal to the ratio of the reflectances, and

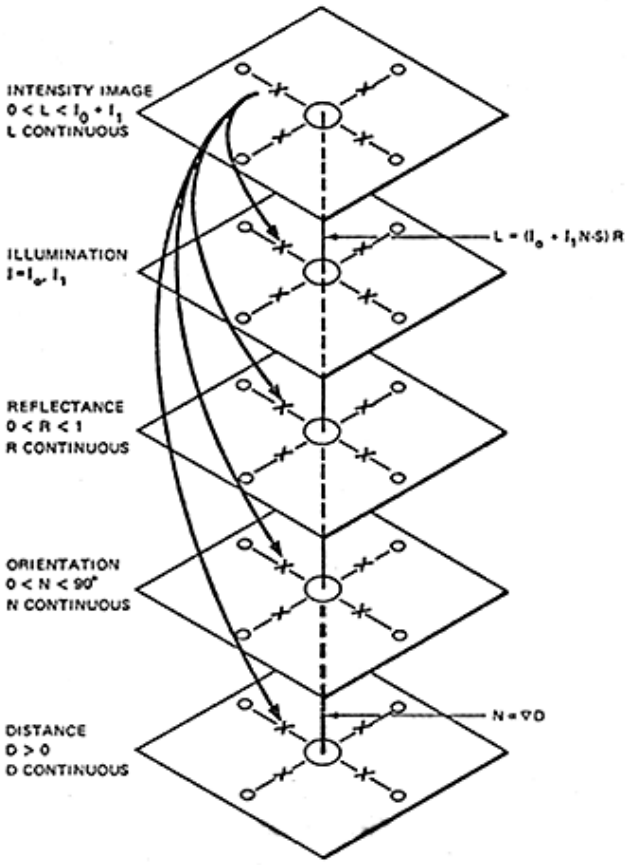


Figure 8 A parallel computational model for recovering intrinsic images

The basic model consists of a stack of registered arrays, representing the original intensity image (top) and the primary intrinsic arrays. Processing is initialized by detecting intensity edges in the original image, interpreting them according to the catalog, and then creating the appropriate edges in the intrinsic images (as implied by the downward sweeping arrows).

Parallel local operations (shown as circles) modify the values in each intrinsic image to make them consistent with the intraimage continuity and limit constraints. Simultaneously, a second set of processes (shown as vertical lines) operates to make the values consistent with interimage photometric constraints. A third set of processes (shown as Xs) operates to insert and delete edge elements, which locally inhibit continuity constraints. The constraint and edge modification processes operate continuously and interact to recover accurate intrinsic scene characteristics and to perfect the initial edge interpretation.

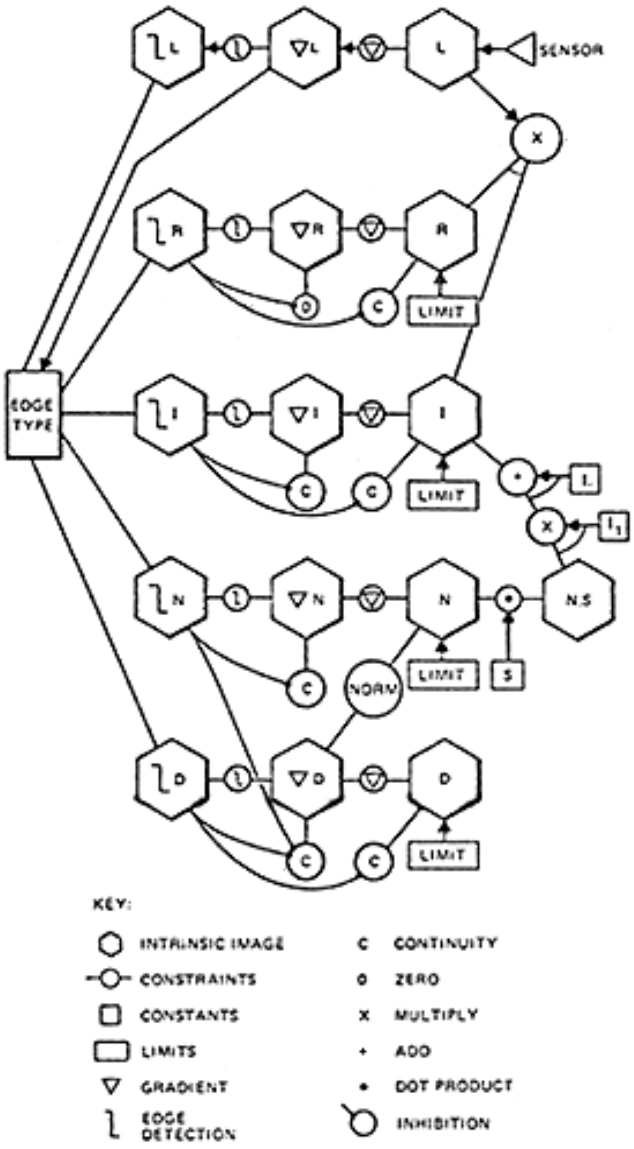


Figure 9 Organization of the computational model

The input intensity inage and primary intrinsic images each have an associated gradient image and an edge image. Additional images represent intermediate results of computation, such as N.S. (All images are shown as hexagons.) The constraints, shown here as circles, are of three varieties: physical limits (e.g., 0 =< R =< 1), local continuity, and interimage consistency of values and edges. Continuity constraints are inhibited across edges. For example, values of illumination gradient are constrained to be continuous, except at illumination edges.

its direction is the same on both sides. These characteristics uniquely identify illuminated reflectance edges, and provide constraints relating intrinsic characteristics on the two sides.

In shadow, it is not possible to locally distinguish a pure reflectance edge from an extremal edge between surfaces of different reflectance, for which the ratio of intensities is also equal to the ratio of reflectances.

The existence of reflectance edges introduces a new type or X-shaped junction, where a shadow edge is cast across a reflectance edge. The detection of an X-junction in the image unambiguously identifies a shadow edge, since the reflectance edge may be easily identified by the ratio test.

An interesting case of surface markings is that of reflectance texture. Texture has certain regular or statistical properties that can be exploited to estimate relative distance and surface orientation. If we can assume statistically uniform density of particular textural features, the apparent density in the image relates distance and inclination of the surface normal to the viewing direction. A second cue is provided by orientation of textural features (for example, reflectance edge elements). As the surface is viewed more obliquely, the orientation distribution of image features tends to cluster about the orientation of the extremal boundary, or of the horizon. These cues are important, since they provide independent information about surface shape, perhaps less precisely and of lower resolution than photometric information, but in areas where photometric information is unavailable (e.g., shadowed regions) or ambiguous (e.g., an illuminated region seen through a hole).

The assumption of smoothness of surfaces can be relaxed by introducing a further edge type, the intersection edge, which represents a discontinuity in surface orientation, such as occurs at a crease or between the faces of a polyhedron. There are two distinct ways an intersection edge can appear in an image, corresponding to whether one or both of the intersecting surfaces are visible. We shall call these subcases "occluding" and "connecting," respectively.

At a connecting intersection edge, only distance is necessarily continuous, since faces can be differently painted, and illuminated by different sources. The strong assumption of continuity of orientation is replaced by the weaker one that the local direction of the surface edge in three dimensions is normal to the orientations of both surfaces forming it. The effect of this constraint is that if one surface orientation is known, the surface edge direction can be inferred from the image, and the other surface orientation is then hinged about that edge, leaving it one degree of freedom. Even when neither orientation is known absolutely, the existence of a connecting edge serves to inhibit application of continuity constraints, and thereby permit more accurate reconstruction of surface shape.

At an occluding intersection edge, nothing is known to be continuous, and the only constraint is on relative distance.

In the image, an illuminated intersecting edge can be distinguished from an extremal edge since the intensity on both sides is varying, but the tangency test fails, and it can be distinguished from a reflectance edge since the ratio of intensities across the edge is not constant. The constraint between surface orientations forming the edge makes it appear likely that a test can also be devised for distinguishing between connecting and occluding intersection edges.

In shadowed regions, intersection edges are only visible when they coincide with reflectance edges, from which they are therefore locally indistinguishable. Creases in a surface are thus invisible in shadows.

When one surface is illuminated and the other shadowed, an intersection edge cannot be locally distinguished from the case of a shadowed object occluding an illuminated one.

Extremal and intersection boundaries together give a great deal of information about surface shape, even in the absence of other evidence, such as shading, or familiarity with the object. Consider, for example, the ability of an artist to convey abstract shape with line drawings. The local inclination of an extremal or intersection boundary to the line of sight is, however, unknown; a given silhouette can be produced in many ways [27]. In the absence of other constraints, the distance continuity constraint will ensure smooth variation of distance at points along the boundary. An additional constraint that could be invoked is to assume that the torsion (and possibly also the curvature) of the boundary space curve is minimal. This will tend to produce planar space curves for boundaries, interpreting a straight line in the image as a straight line in space, or an ellipse in the image as a circle in space. The assumption of planarity is often very reasonable: it is the condition used by Marr to interpret silhouettes as generalized cylinders [27].

The assumption of known illumination can be relaxed in various ways. Suppose we have the same "sun and sky" model of light sources, but do not have prior knowledge of I0, I1 and S. In general, we cannot determine the flux densities absolutely, since they trade off against reflectance. We may, however, assign an arbitrary reflectance to one surface (for example, assume the brightest region is white, $R = 1.0$), and then determine other reflectances and the flux densities relative to it. The initial assignment may need to be changed if it forces the reflectance of any region to exceed the limits, $0.0 < R < 1.0$.

The parameters of illumination, flux densities I0 and I1, and unit vector S, can be determined by assuming reflectance and exploiting a variation of the tangency test. If we have an illuminated extremal boundary, Equation (5) gives a linear equation in the parameters for each point on the edge. The equations for any four points can be solved simultaneously to yield all the parameters. The remaining points on the boundary can be used with the now-known illumination to verify the assumption of an extremal boundary. An independent check on the ratio of I0 to I1 can be made at any shadow edge where surface orientation is known. This method of solving for illumination parameters can be extended to multiple point sources by merely increasing the number of points on the boundary used to provide equations. Care is required with multiple sources because it is necessary to know which points are illuminated by which sources. The method works for a centrally projected image, but not for an orthogonally projected one, since, in the latter case, all the known surface normals are coplanar.

Even modelling illumination as a set of point sources does not capture all the subtleties of real-world lighting, which include extended sources, possibly not distant, and secondary illumination. Extended sources cause shadows to have fuzzy edges, and close sources cause significant gradients in flux density and direction. Secondary illumination causes local perturbations of illumination that can even dominate primary illumination, under some circumstances (e.g., light scattered into shadow regions). All these effects make exact modelling of illumination difficult, and hence cast suspicion on a recovery method that requires such models.

In the absence of precise illumination models that specify magnitude and direction distributions of flux density everywhere, accurate point-wise estimation of reflectance and surface orientation from photometric equations alone is not possible. It should still be possible, however, to exploit basic photometric constraints, such as local continuity of illumination, along with other imaging constraints and domain assumptions, to effect recovery within our general paradigm. As an example, we might still be able to find and classify edges in the intensity image: reflectance edges still have constant intensity ratios (less than 30:1) across them, shadow edges can be fuzzy with high intensity ratios, occlusion and intersection edges are generally sharp without constant ratios. The occlusion and intersection edges, together with reflectance texture gradient and continuity assumptions, should still provide a reasonable initial shape estimate. The resulting knowledge of surface continuity, the identified reflectance edges, and the assumption of reflectance constancy enable recovery of relative reflectance, and hence relative total incident flux density. The ability to determine continuity of illumination and to discriminate reflectance edges from other types thus allows us to generalize Horn's lightness determination process [18] to scenes with relief, occlusion, and shadows.

Having now made initial estimates of intrinsic characteristics, it may be possible to refine them using local illumination models and photometric knowledge. It may be possible, using assumptions of local monotonicity of illumination, to decide within regions whether the surface is planar, curved in some direction, or whether it inflects.

Even with all the extensions that we have so far discussed, our scene domain is still much less complex than the real world, in which one finds specularity, transparency, luster, visible light sources, three-dimensional texture, and sundry other complications. Although, at first sight, these phenomena make recovery seem much harder, they also represent additional intrinsic characteristics for more completely describing the scene, and potentially rich sources of information for forming the description. There are also many well-known sources of information about the scene,

that make use of multiple images, including stereo, motion parallax, and color. We believe that the framework we have put forward can be extended to accommodate these additional sources.

At this point, it is not clear whether adding complexity to our domain will lead to fewer ambiguities, or more. So far, however, we have seen no reason to rule out the feasibility of recovering intrinsic scene descriptions in realworld situations.

VII DISCUSSION

The concept of intrinsic images clarifies a number of issues in vision and generalizes and unifies many previously disjoint computational techniques. In this section, we will discuss some implications of our work, in the contexts of system organization, image analysis, scene analysis, and human vision.

A. System Organization

The proper organization of a visual system has been the subject of considerable debate. Issues raised include the controversy over whether processing should be data-driven or goal-driven, serial or parallel, the level at which the transition from iconic to symbolic representation should occur, and whether knowledge should be primarily domain-independent or domain-specific [4]. These issues have a natural resolution in the context of a system organized around the recovery of intrinsic images, as in Figure 10.

The recovery process we have outlined is primarily data-driven and dependent only on general domain assumptions. Subsequent goal-oriented or domain-specific processes (ranging from segmentation to object recognition) may then operate on information in the intrinsic images.

Intrinsic images appear to be a natural interface between iconic and symbolic representations. The recovery process seems inherently iconic, and suited to implementation as an array of parallel processes attempting to satisfy local constraints. The resulting information is iconic, invariant, and at an appropriate level for extracting symbolic scene descriptions. Conversely, symbolic information from higher levels (e.g., the size or color of known objects) can be converted to iconic form (e.g., estimates of distance or reflectivity) and used to refine or influence the recovery process. Symbolic information clearly plays an important role in the perception of the three-dimensional world.

Multilevel, parallel constraint satisfaction is an appealing way to organize a visual system because it facilitates incremental addition of knowledge and avoids many arbitrary problems of control and sequencing. Parallel implementations have been used previously, but only at a single level, for recovering lightness [18], depth [28], and shape [35]. Each of these processes is concerned with recovering one of our intrinsic images, under specialized assumptions equivalent to assuming values for the other images. In this paper, we have suggested how they might be coherently integrated.

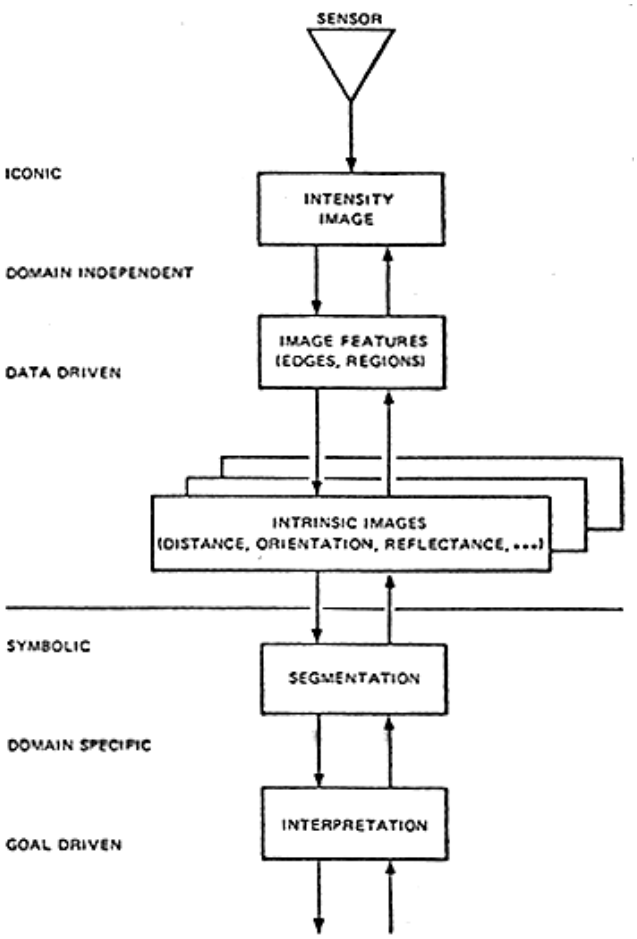


Figure 10 Organization of a visual system

Issues of stability and speed of convergence always arise in iterative approaches to constraint satisraction. Heuristic "relaxation" schemes (e.g., [32]), in which "probabilities" are adjusted, often use ad hoc updating rules for which convergence is difficult to obtain and prove. By contrast, the system we have described uses iterative relaxation methods to solve a set of equations and inequalities. The mathematical principles of iterative equation solving are well understood [2] and should apply to our system, at least, for a fixed set of edges. Insofar as local edge modifications have only local consequences, operations such as gap filling should affect convergence to only a minor extent.

Speed of convergence is sometimes raised as an objection to the use of cooperative processes in practical visual systems; it is argued that such processes would be too slow in converging to explain the apparent ability of humans to interpret an image in a fraction

of a second. This objection can be countered in several ways: first, full convergence may not be necessary to achieve acceptable accuracy; second, information propagation may be predominantly local, influenced primarily by nearby edges; third, there are ways of speeding up long-range propagation -- for example, using hierarchies of resolution [15].

B. Image Analysis

Image analysis is usually considered to be the generation of a two-dimensional description of the image, such as a line drawing, which may subsequently be interpreted. We believe it is important to take a more liberal view, and include some consideration of the three-dimensional meaning of image features in the earliest descriptions.

A topic often debated is segmentation -- the process of partitioning an image into semantically interpretable regions. Fischler [9] and others have raised a number of critical questions: is segmentation meaningful in a domain-independent sense, is it possible, and how should it be done?

Partitioning of an arbitrary intensity image into regions denoting objects is an illusory goal, unless the notion of what constitutes an object is precisely defined. Since objects are often collections of pieces whose association must be learned, general object segmentation seems impossible in principle. It seems more reasonable, on the other hand, to partition an image into regions corresponding to smooth surfaces of uniform reflectance. This is often the implicit goal of programs that attempt to partition an image into regions of uniform intensity. Unfortunately, intensity does not correspond directly to surface characteristics. There is no way of determining whether merging two regions is meaningful, and consequently there is no reliable criterion for terminating the merging process. Segmentation based on intrinsic images avoids these difficulties.

Another elusive goal, the extraction of a perfect line drawing from an intensity image, is also impossible in principle, for similar reasons: the physical significance of boundaries does not correlate well with the magnitude of intensity changes. Surface boundaries can be hard, and, in some places, impossible, to detect; shadows and texture contribute edge points in abundance, which, in this context, are noise. To attain a line drawing depicting surface boundaries, we must take into account the physical significance of intensity discontinuities. It is quite clear from depth and orientation intrinsic images where edges are necessary for consistency and surface integrity. A perfect line drawing could be regarded as one of the products of the process of recovering intrinsic characteristics.

From this point of view, all attempts to develop more sophisticated techniques for extracting line drawings from intensity images appear inherently limited. Recently, relaxation enhancement techniques for refining the imperfect results of an edge detector have attracted considerable interest [32 and 15]. These techniques manipulate edge confidences according to the confidences of nearby points, iterating until equilibrium is achieved. This approach is really attempting to introduce and exploit the concept of edge continuity, and does lead to modest improvements. It does not, however, exploit the continuity of surfaces, nor ideas of edge type, and consequently produces curious results on occasion. Moreover, as we noted earlier, convergence for ad hoc updating rules is difficult to prove.

The major problem with all the image analysis techniques we have mentioned is that they are based on characteristics of the image without regard to how the image was formed. Horn at MIT for some time has urged the importance of understanding the physical basis of image intersity variations [20]. His techniques for determining surface lightness [18] and shape from shading [19] have had an obvious influence on our own thinking. To achieve a precise understanding of these phenomena, Horn considered each in isolation, in an appropriately simplified domain: a plane surface bearing regions of different reflectance lit by smoothly varying illumination for lightress, and a simple smoothly curved surface with uniform reflectance and illumination for shading. These domains are, however, incompatible, and the techniques are not directly applicable in domains where variations in reflectance, illumination, and shape may be discontinuous and confounded. We have attempted to make explicit the constraints and assumptions underlying such recovery techniques, so that they may be integrated and used in more general scenes.

The work most closely related to our own is that of Marr [29], who has described a layered organization for a general vision system. The first layer extracts a symbolic description of the edges and shading in an intensity image, know as the "Primal Sketch." These features are intended to be used by a variety of pro-cesses at the next layer (e.g., stereo correla-tion, texture gradient) to derive the three-dimensional surface structure of the image. The resulting level of description is analogous to our orientation and distance images, and is called the "2.5D sketch." Our general philosophy is similar to Marr's, but differs in emphasis, being somewhat complementary. We are all interested in understarding the organization of visual systems, in terms of levels of representation and information flow. Marr, however, has concentrated primarily on understanding the nature of individual cues, such as stereopsis and texture, while we have concentrated primarily on understanding the integration of multiple cues. We strongly believe that interaction of different kinds of constraints plays a vital role in unscrambling information about intrinsic scene characteristics.

A particular point of departure is Marr's reliance on two-dimensional image description and grouping techniques to perfect the primal sketch before undertaking any higher-level processing. By contrast, we attempt to immediately assign three-dimensional interpretations to intensity edges to initialize processing at the level of intrinsic images, and we maintain the relationship between intensities and interpretations as tightly as possible. In our view, perfecting the intrinsic images should be the objective of early visual processing; edges at the level of the primal sketch are the consequence of achieving a consistent three-dimensional interpretation. We shall discuss consequences of these differing organizations with reference to human vision shortly.

C. Scene Analysis

Scene analysis is concerned with interpreting images in three dimensions, in terms of surfaces, volumes, objects, and their interrelationships. The earliest work, on polyhedral scenes, involved extracting line drawings and then interpreting them using geometric object prototypes [31, 8]. A complementary approach analyzed scenes of simple curved objects by partitioning the image into regions of homogeneous intensity and then interpreting them using relational models of object appearances [5, 3]. Both these early approaches were limited by the unreliability of extracting image descriptions, as discussed in the preceding section, and by the lack of generality of the object prototypes used. It was soon discovered that to extract image descriptions reliably required exploiting knowledge of the scene and image formation process. Accordingly, attempts were made to integrate segmentation and interpretation. The general approach was to assign sets of alternative interpretations to regions of uniform intensity (or color), and then alternately merge regions with compatible interpretations and refine the interpretation sets. The process terminates with a small number of regions with disjoint (and hopefully unique) interpretations. Yakimovsky and Feldman [36] used Bayesian statistics for assigning interpretations and guiding a search for the set of regions and interpretations with the highest joint likelihood. Tenenbaum and Barrow (IGS [33]) used an inference procedure, similar to Waltz's filtering [34], for eliminating inconsistent interpretations. These systems performed creditably upon complex images and have been applied in a variety of scene domains. They are not, however, suitable as models of general-purpose vision because they are applicable only when all objects are known and can be distinguished on the basis of local evidence or region attributes. Unknown objects not only cannot be recognized; they cannot even be described.

What seems needed in a general-purpose vision system are more concise and more general descriptions of the scene, at a lower level than objects [4 and 38]. For example, once the scene has been described at the level of intrinsic surface characteristics, surfaces and volumes can be found and objects can then readily be identified. There is still the need to guide formation of lower level descriptions by the context of higher level ones, but now the gaps between levels are much smaller and easier to bridge.

Huffman [21], Clowes [6], and Waltz [34] demonstrated the possibility of interpreting line drawings of polyhedral scenes without the need for specific object prototypes. Their level of description involved a small set of linear scene features (convex edge, concave edge, shadow edge, crack) and a set of corner features, where such edges meet. These scene features were studied systematically to derive a catalog of corresponding image features and their alternative interpretations. Interpretation of the line drawing involved a combinatorial search for a compatible set of junction labels. Waltz, in particular, identified eleven types of linear feature, and three cases of illumination for the surfaces on each side. The resulting catalog of junction interpretations contained many thousands of entries. To avoid combinatorial explosion in determining the correct interpretations, Waltz used a pseudo-parallel local filtering paradigm that eliminated junction interpretations incompatible with any possible interpretation of a neighboring junction.

While we also create and use a catalog, the whole emphasis of our work is different. We have attempted to catalog the appearances of edges in grey-scale images, for a much wider class of objects, and have described them in a way that results in a much more parsimonious catalog. Instead of interpreting an ideal line drawing, we, in a sense, are attempting simultaneously to derive the drawing and to interpret it, using the interpretation to guide the derivation. In contrast to the junctions in line drawings, many gray-scale image features can be uniquely interpreted using intensity information and physical corstraints. We are thus able to avoid combinatorial search and "solve" directly for consistent interpretations of remaining features. Our solution has a definite iconic flavor, whereas Waltz’s has largely a symbolic one.

Mackworth's approach to interpreting line drawings [24] is somewhat closer to our point of view. He attempts to interpret edges, rather than junctions, with only two basic interpretations (Connect and Occluding); he models surfaces by representing their plane orientations explicitly; and he tries to solve constraints to find orientations that are consistent with the edge interpretations. The use of explicit surface orientation enables Mackworth to reject certain interpretations with impossible geometry, which are accepted by Waltz. Since he does not, however, make explicit use of distance information, there are still some geometrically impossible interpretations that Mackworth will accept. Moreover, since he does not use photometry, his solutions are necessarily ambiguous: Horn has demonstrated [20] that when photometric information is combined with geometry, the orientations of surfaces forming a trihedral corner may be uniquely determined. One fundamental point to be noted is that intrinsic characteristics provide a concise description of the scene that enables rejection of physically impossible interpretations.

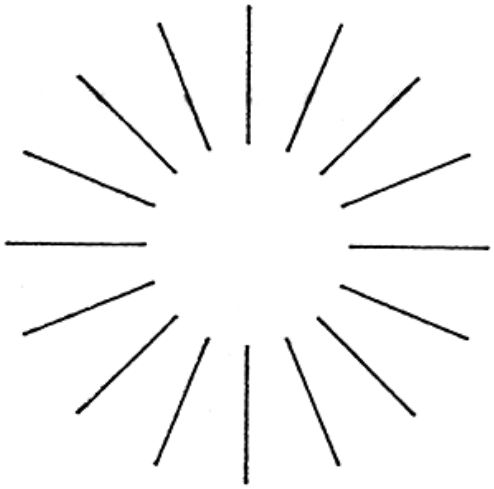
Intermediate levels of representation have played an increasingly important role in recent scene analysis research. Agin and Binford [1] proposed a specific representation of three-dimensional surfaces, as generalized cylinders, and described a system for extracting such representations using a laser rangefinder. Marr and Nishihara [29] have described techniques for inferring a generalized cylinder representation from line drawings, and for matching geometrically to object prototypes in recognition. Cylindrical volume representations are being used in the VISIONS system, under development by Riseman et al. [16]. This system also includes explicit representation of surfaces, which are inferred from a two-dimensional image description using higher-level cues, such as linear perspective and the shape of known objects. Symbolic representations at the level of surfaces and volumes should be easier to derive from intrinsic images than from intensity images, line drawings, or even from noisy rangefinder data.

D. Human Vision

In this paper, we have been concerned with the computational nature of vision tasks, largely independent of implementation in biological or artificial systems. This orientation addresses questions of competence: what information is needed to accomplish a task, is it recoverable from the sensed image, that additional constraints, in the form of assumptions about the world, are needed to effect the recovery?

Psychologists have been asking similar questions fron their own viewpoint for many years. For example, there has been a long-standing debate, dating back at least to Helmholtz, concerning how, and under what circumstances, it is possible to independently estimate illumination and reflectance. Recent participants include Land, with his Retinex theory of color perception [23], and Gilchrist, who has identified a number of ways in which intensity edges may be classified (e.g., the 30:1 ratio, intersecting reflectance and illumination edges) [12].

From such work, a number of theories have been proposed to explain human abilities to estimate various scene characteristics. No one, however, has yet proposed a comprehensive model integrating all known abilities. While we have no intention of putting forward our model as a complete explanation of human vision, we think that the recovery of intrinsic characteristics is a plausible role for early stages of visual processing in humans.



Figuree 11 A subjective contour

This hypothesis would appear to explain, at least at a superficial level, many well-known psychological phenomena. The constancy phenomena are the obvious examples, but there are others. Consider, for example, the phenomenon of subjective contours, such as appear in Figure 11. Marr suggests [26] that such contours result from grouping place tokens corresponding to line endings in the primal sketch, and further suggests a "least-commitment" clustering algorithm to control the combinatorics of grouping. We suggest, as an alternative explanation, that the abrupt line endings are locally interpreted three dimensionally as evidence of occlusion, causing discontinuity edges to be established in the distance image. The subjective contours then result fron making the distance image consistent with these boundary conditions and assumptions of continuity of surfaces and surface boundaries: they are primarily contours of distance, rather than intensity. The net result is the interpretation of the image as a disk occluding a set of radiating lines on a more distant surface.

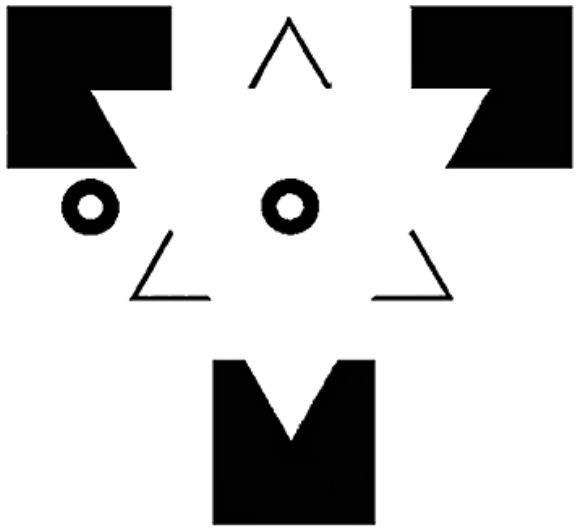


Figure 12 Subjective depth (Coren [7])

There is considerable evidence to support the hypothesis that subjective contours are closely correlated with judgements of apparent depth. Coren [7] reports a very interesting demonstration. In Figure 12, the two circles subtend the same visual angle; however, the apparent elevation of the subjective triangle causes the circle within it to be perceived as smaller, consistent with the hypothesis that it is nearer. Surfaces perceived when viewing Julesz stereograms [22] have edges that are purely subjective. There are no distinguishing cues in the originating intensity images: the edges result solely from the discontinuity in disparity observed in a stereo presentation. Hochberg [17] has investigated the subjective contours produced by shadow cues to depth, seen in figures like Figure 13. Most observers report that Figure 13a is perceived as a single entity in relief, while Figure 13b is not. The figures are essentially equivalent as two-dimensional configurations of lines. The difference between the figures is that in b, the lines are not consistent with the shadows of a single solid entity cast by a directional light source.

We can generalize our argument that subjective contours arise as a consequence of three-dimensional organization to other phenomena of perceptual organization, such as the Gestalt laws. For example, the law of continuity follows directly from assumptions of continuity of surfaces and boundaries, and

the law of closure follows from integrity of surfaces.

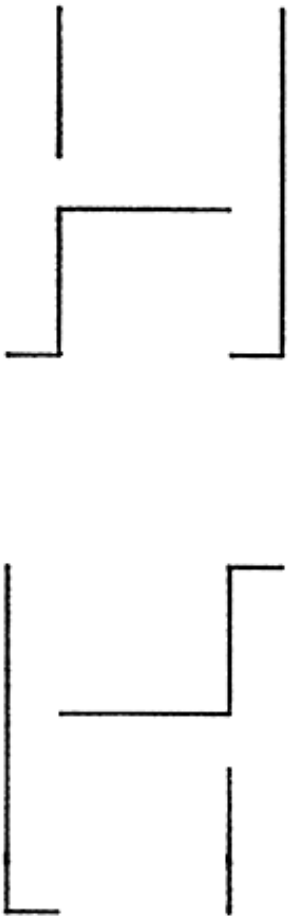


Figure 13 Subjective figures (Hochberg [17])

A system that is attempting to form a consistent interpretation of an image as a three-dimensional scene will, in principle, do better than one that is attempting to describe it as a two-dimensional pattern. The organization of a chaotic collection of image features may become clear when they are considered as projections of three-dimensional features, and the corresponding constraints are brought into play.

Gregory has emphasized the importance of three-dimensional interpretations and has suggested that many illusions result from attempting to form inappropriate three-dimensional interpretations of two-dimensional patterns [13 and 14]. He also suggests that certain other illusions, such as those involving the Ames room, result from applying incorrect assumptions about the nature of the scene. The distress associated with viewing ambiguous figures, such as the well-known “impossible triangle” and “devil's pitchfork,” arises because of the impossibility of making local evidence consistent with assumptions of surface continuity and integrity.

Recent experiments by Gilchrist [11] demonstrate that judgements of the primary intrinsic characteristics are tightly integrated in the human visual system. In one experiment, the apparent position of a surface is manipulated using interposition cues, so that it is perceived as being either in an area of strong illumination or in one of dim illumination. Since the absolute reflected flux from the surface remains constant, the observer perceives a dramatic change in apparent reflectance, virtually from black to white. In a second experiment, the apparent orientation of a surface is changed by means of perspective cues. The apparent reflectance again changes dramatically, depending upon whether the surface is perceived as facing towards or away from the light source.

We do not present our model of the recovery of intrinsic scene characteristics as a comprehensive explanation of these psychological phenomena. We feel, however, that it may provide a useful viewpoint for considering at least some of them.

VIII CONCLUSION

The key ideas we have attempted to convey in this paper are:

* A robust visual system should be organized around a noncognitive, nonpurposive level of processing that attempts to recover an intrinsic description of the scene.
* The output ot this process is a set of registered "intrinsic images" that give values for such characteristics as reflectance, orientation, distance, and incident illumination, for every visible point in the scene.
* The information provided by intrinsic images greatly facilitates many higher-level perceptual operations, ranging from segmentation to object recognition, that have so far proved difficult to implement reliably.
* The recovery of intrinsic scene characteristics is a plausible role for the early stages of human visual processing.
* Investigation of low-level processing should focus on what type of information is sought, and how it might be obtained from the image. For example, the design of edge detectors should be based on the physical meaning of the type of edge sought, rather than on some abstract model of an intensity discontinuity.

We have outlined a possible model of the recovery process, demonstrated its feasibility for a simple domain, and argued that it can be extended in a straightforward way towards real-world scenes. Key ideas in the recovery process are:

* Information about the intrinsic characteristics is confounded in the intensities of the input image. Therefore, recovery depends on exploiting assumptions and constraints from the physical nature of imaging and the world.
* Interactions and ambiguities prohibit independent recovery of the intrinsic

characteristics; the recovery process must determine their values simultaneously in achieving consisteney with the constraints.

* Interpretation of boundaries plays a key role in recovery; they provide information about which characteristics are continuous and which discontinuous at each point in the image, and they provide explicit boundary conditions for the solution.
* The nature of the solution, involving a large number of interacting local constraints, suggests implementation of the recovery process by an array of local parallel processes that achieve consistency by a sequence of successive approximations.

Our model for recovering intrinsic characteristics is at a formative stage. Important details still to be finalized include the appropriate intrinsic images, constraints, constraint representations, and the paths of information flow relating them. Nevertheless, the ideas we have put forth in this paper have already clarified many issues for us and suggested many exciting new prospects. They also raise many questions to be answered by future research, the most important of which are "Can it work in the real world?" and "Do people see that way?" To the extent that our model corresponds to the human visual system, valuable insights may be gained through collaboration between computer and vision.

IX ACKNOWLEDGEMENTS

This work was supported by a grant from the National Science Foundation.

Edward Riseman's persistent exhortations provided the necessary motivation to commit our ideas to paper, and his valuable editorial comments improved their presentation.

Bill Park's artistic hand penned the drawings of intrinsic images.

REFERENCES

1. Agin, G. J. and Binford, T.O. Computer description of curved objects. Proc. 3rd. International Joint Conference on Artificial Intelligence, Stanford University, Stanford, 1973.

2. Allen, D. N. de G. Relaxation Methods in Engineering and Science. McGraw-Hill, New York, 1954.

3. Barrow, H. G., and Popplestone, R. J. Relational descriptions in picture processing. in Machine Intelligence, Vol. 6, B. Meltzer and D. Michie (Eds.), Edinburgh University Press, Edinburgh, Scotland, 1971, pp 377-396.

4. Barrow, H. G., and Tenenbaum, J. M. Representation and use of knowledge in vision. Tech. Note 108, Artificial Intelligence Center, SRI International, Menlo Park, CA, 1975.

5. Brice, C., and Fennema, C. Scene analysis using regions. Artificial Intelligence, 1, No. 3, 1970, 205-226.

6. Clowes, H. B. On seeing things. Artificial Intelligence, 2, No. 1, 1971, 79-112.

7. Coren, S. Subjective contour and apparent depth. Psychological Review, 79, 1972, 359.

8. Falk, G. Interpretation of imperfect line data as a three-dimensional scene. Artificial Intelligence, 4, No. 2, 1972, 101-144.

9. Fischler, M. A. On the representation of natural scenes, in Computer Vision Systems, A. Hanson and E. Riseman (Eds.), Academic Press, New York, 1978.

10. Garvey, T. D. Perceptual strategies for purposive vision. Tech Note 117, Artiticial Intelligence Center, SRI International, Menlo Park, CA, 1976.

11. Gilchrist, A. L. Perceived lightness depends on perceived spatial arrrangement. Science, 195, 1977, 185-187.

12. Gilchrist, A. L. Private communication.

13. Gregory, R. L., Eye and Brain. Weidenfeld and Nicholson, London, 1956.

14. Gregory, R. L. The Intelligent Eye. McGraw-Hill, New York, 1970.

15. Hanson, A., and Riseman, E, Segmentation of natural scenes, in Computer Vision Systems, A. Hanson and E. Riseman (eds.), Academic Press, New York, 1978.

15. Hanson, A., and Riseman, E. VISIONS. A computer system for interpreting scenes, in Computer Vision Systems, A. Panson and E. Riseman (Eds.), Academic Press, New York, 1978.

17. Hocherg, J. In the mind's eye, in Contemporary Theory and Research in Visual Perception, R. N. Haber (Ed.), Holt, Rinehart and Winston, New York, 1960.

18. Horn, B. K. P. Determining Lightness from an image. Computer Graphics and Image Processing, 3, 1974, 277-299.

19. Horn, B. K. P. Obtaining shape from shading information. in The Psychology of Computer Vision, P. H. Winston (Ed.), McGraw-Hill, New York, 1975.

20. Horn, B. K. P. Understanding image intensities. Artificial Intelligence, 8, No. 2, 1977, 201-231.

21. Huffman, D. A. Impossible objects as nonsense sentences. in Machine Intelligence, Vol. 6, B. Meltzer and D. Michle (Eds.), Edinburgh University Press, Edinburgh, 1971, pp 295-323.

22. Julesz, B. Foundations of Cyclopean Perception. University of Chicago Press, Chicago, 1971.

23. Land, E. H. The Retinex theory of color vision. Scientific American, 237, No. 6, Dec. 1977, 108-128.

24. Mackworth, A. K. Interpreting pictures of polyhedral scenes. Artificial Intelligence, 4, 1973, 121-138.

25. Macleod, I. D. G. A study in automatic photo-interpretation. Ph. D. thesis, Department of Engineering Physics, Australian National University, Canberra, 1970.

26. Marr, D. Early processing of visual information. AI Memo-340, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1976.

27. Marr, D. Analysis of occluding contour. Proc. Roy. Soc. Lond. B, 197, 1977, 441-475.

28. Marr, D., and Poggio, T. Cooperative computation of stereo disparity. Science, 194, 1977, 283-287 .

29. Marr, D. Representing visual information. in Computer Vision Syetems, A. Hanson and E. Riseman (Eds.), Academic Press, New York, 1978.

30. Nitzan, D., Brain, A. E., and Duda, R. O. The measurement and use of registered reflectance and range data in scene analysis. Proc. IEEE, 65, No. 2, 1977, 206-220.

31. Roberts, L. G. Machine perception of three-dimensional solids. in Optical and Electro-optical Information Processing, J. T. Tippett et al. (Eds.), MIT Press Cambridge, Mass., 1965.

32. Rosenfeld, A., Hummel, R. A., and Zucker, S. W. Scene labeling by relaxation operations", IEEE Transactions on Systems, Man and Cybernetics, SMC-5, 1976, 420-433.

33. Tenenbaum, J. M., and Barrow, H. G. Experiments in interpretation-guided segmentation. Artificial Intelligence, 8, No. 3, 1977, 241-274.

34. Waltz, D. L. Generating semantic descriptions from drawings of scenes with shadows. Tech. Rept. AI-TR-271, Artificial Intelligence Lab., MIT, Cambridge, Mass., Nov. 1972.

35. Woodham, R. J. A cooperative algorithm for determining surface orientation from a single view. Proc. 5th. International Joint Conference on Artificial Intelligence, MIT, Cambridge, Mass., 1977.

36. Yakinovsky, Y. Y., and Feldman, J. A semantics-based decision theoretic region analyzer. Proc. 3rd. International Joint Conference on Artificial Intelligence, Stanford Univ., Stanford, 1973.

37. Yonas, A. Private communication, Institute of Child Development, University of Minnesota, 1977.

38. Zucker, S., Rosenfeld, A., and Davis, L. General purpose models: Expectations about the unexpected. Proc. 4th. International Joint Conference on Artificial Intelligence, Tbilisi, Georgia, USSR, 1975.