

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 nonpurposeful 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 considerable 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 approaches; those 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 changes 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.

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 horizontal 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 of 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, specularly, 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)

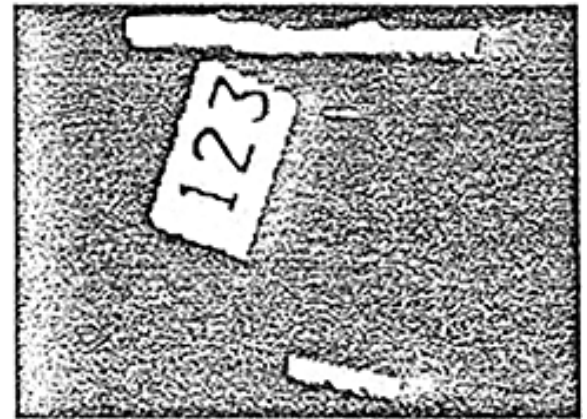
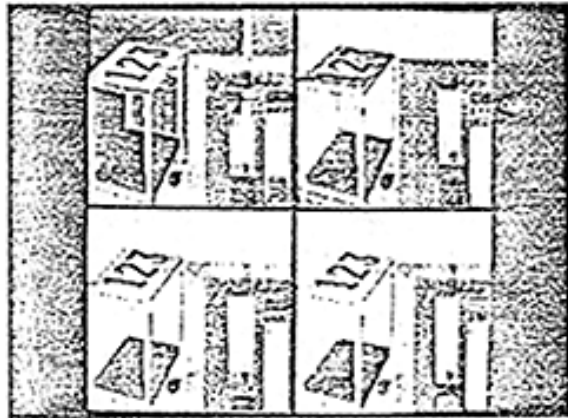
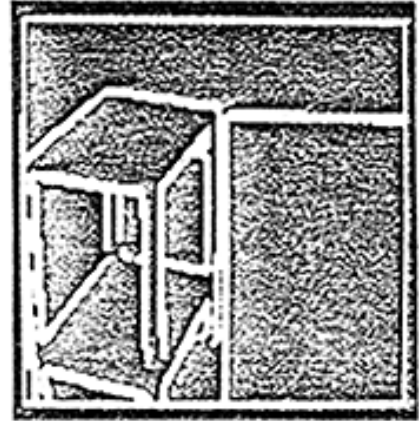
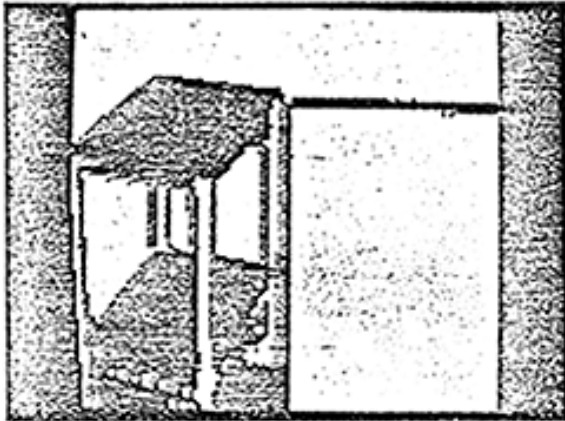
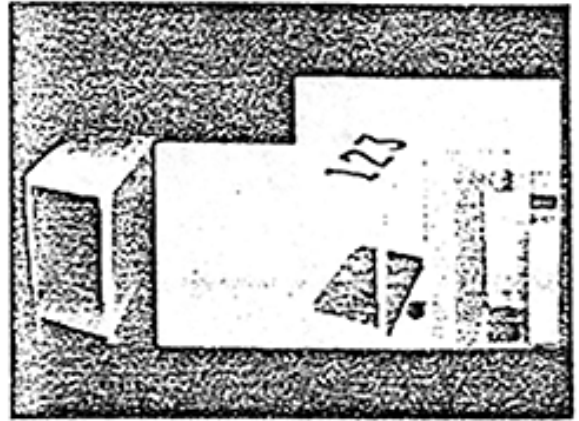
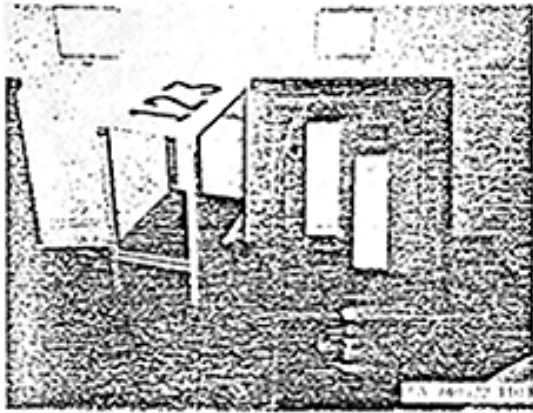


(c) FLAX (X 1000)



(d) WALLFLOWER (X 1800)

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



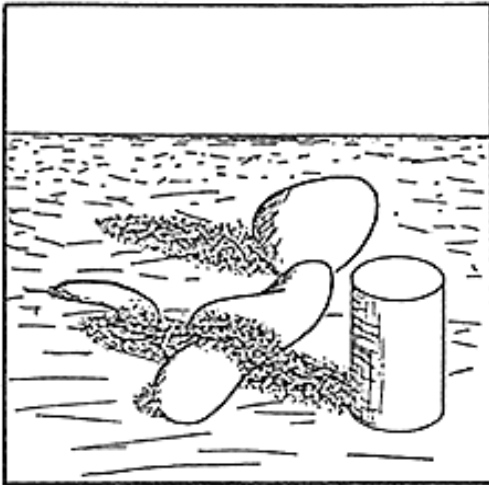
(a) THRESHOLDING REFLECTANCE

(b) CORRECTED VIEW OF CART TOP

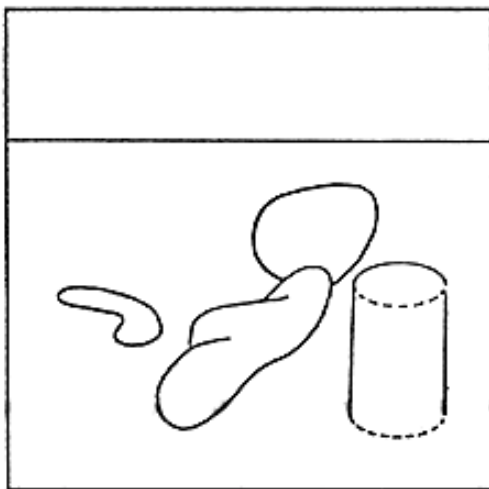
Figure 2 Experiments with a laser rangefinder

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

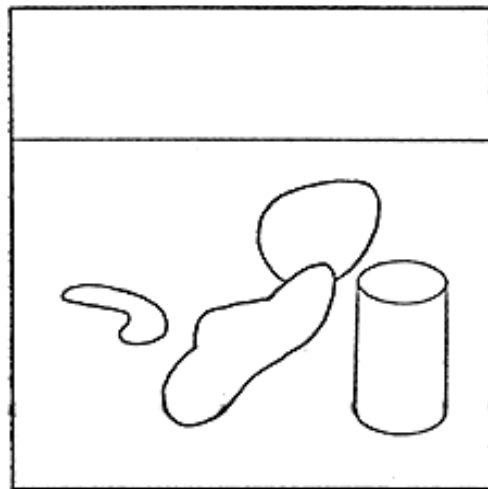
The images are depicted as line drawings, but, in fact, would contain values at every point. The solid lines in the intrinsic images represent discontinuities in the scene characteristic; the dashed lines represent discontinuities in its derivative.



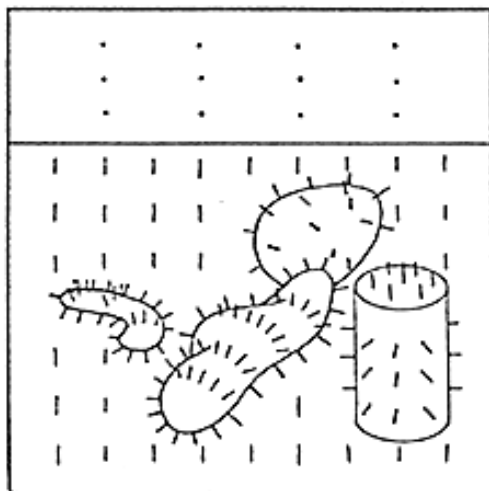
(a) ORIGINAL SCENE



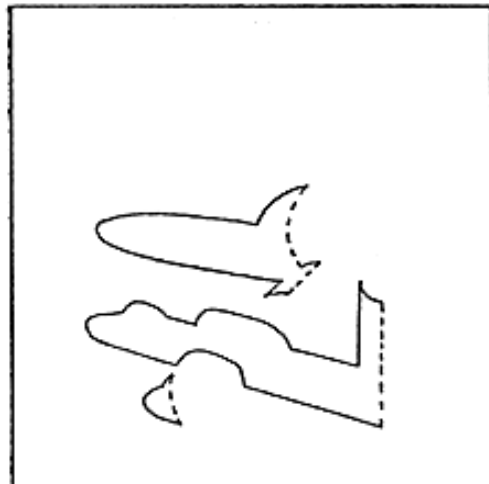
(b) DISTANCE



(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 normal 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.

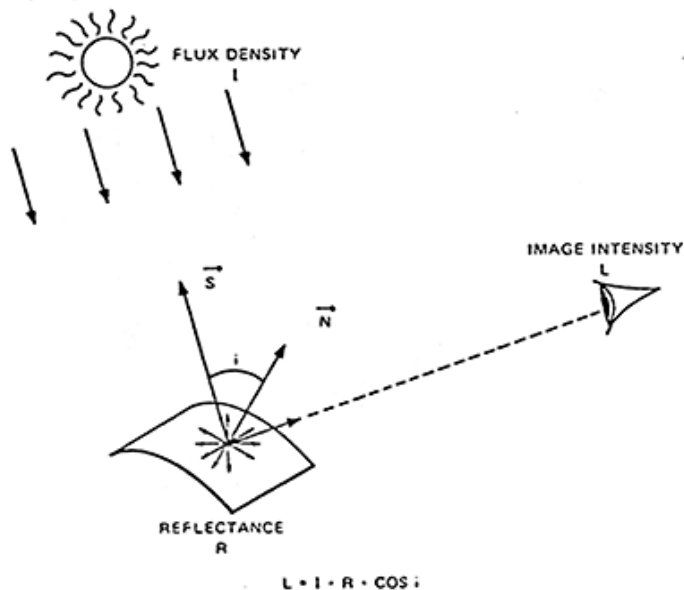


Figure 4 An ideally diffusing surface

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

$$L = I \cdot R \cdot \cos i \quad (1)$$

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 characteristics are continuous through shadows. In man-made environments, straight edges frequently 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

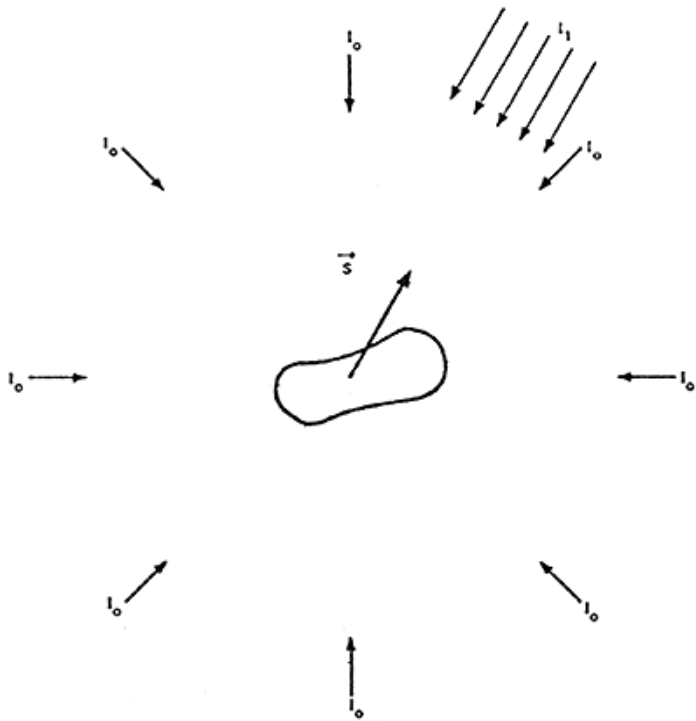


Figure 5 Sun and sky illumination model

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,

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

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 I_0 , the image intensity is given by

$$L = I_0 * R \quad (3)$$

When a surface element is illuminated by a point source, such that the flux density is I_1 , from a direction specified by the unit

vector, S , the incident flux density at the surface is $I_1 * N.S$, where N is the unit normal to the surface, and $.$ is the vector dot product. Thus,

$$L = I_1 * N.S * R \quad (4)$$

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

$$L = (I_0 + I_1 + N.S) * R \quad (5)$$

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 , I_0 , and I_1 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.