ESSAYS IN COGNITIVE PSYCHOLOGY

VISUAL PROCESSING:

COMPUTATIONAL, PSYCHOPHYSICAL AND COGNITIVE RESEARCH

Roger Watt

Visual Processing: Computational, Psychophysical, and Cognitive Research

R. J. Watt

MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge, U.K.

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Hove and London (UK) Hillsdale (USA)

Lawrence Erlbaum Associates Ltd., Publishers
27 Palmeira Mansions
Church Road
Hove
East Sussex, BN3 2FA
U.K.

British Library Cataloguing in Publication Data

Watt, R.J.
Visual processing : computational,
psychophysical and cognitive research.
1. Visual perception
I. Title
152.1'4

ISBN 0-86377-081-9

Typeset by Latimer Trend & Company Ltd, Plymouth
Printed and bound by A. Wheaton & Co. Ltd., Exeter

Acknowledgements

Many people have contributed to this essay. Some did so unwittingly by making opportune comments out of context: Issues of representation are similar whether it is visual space or creative thoughts or emotions that are to be described by the representation. Others have more deliberately challenged the crude ideas that eventually led to the arguments of this essay, and I would especially like to record my gratitude to D. Andrews, D. Foster, S. Laughlin, and M. Morgan.

I should like to thank A. Baddeley, V. Bruce, B. Craven, P. Johnson-Laird, B. Moulden, D. Osorio, T. Valentine, and A. Wilkins for their critical comments on the presentation of my ideas. Valda Jones typed, retyped, reretyped this until she was blue in the face. I shall be eternally grateful.

Preface: Scope and Purpose of this Essay

Vision seems effortless and simple to us, the users. We can sense the shape, structure, and spatial layout of a large number of remote objects very rapidly and usually with a high level of accuracy and stability. Our only conscious actions are to point the eyes in the right direction, focus them, and apply the ephemeral commodity of attention to items of especial interest or importance. Even these actions seem effortless in normal use: when scanning a road before stepping off the pavement; or when reading text, examining graphs, watching the television.

It has, however, been found to be extremely difficult to make machinery that can form images from light already available in the environment, and then interpret such images in terms of the three-dimensional scene. The failure to build a general-purpose vision system has many roots, some of which are particularly instructive with important lessons on how not to think about vision. For instance, it is quite usual to start with two cameras and then to recover range (depth) data from the geometry of stereopsis. After all, we have two eyes in our head, and obviously need both for the vivid stereoscopic sensation of depth. But is depth perception the real reason why we have two eyes? Perhaps we have two eyes to allow for mechanical faults, and indeed more than 10% of the population of Western countries (and much higher proportions in Third World countries) have a fault in one eye that renders it useless. This is the condition known as amblyopia, and the perception of distance by amblyopes is not impaired to any significant extent. The lesson for engineers is that they should avoid unnecessary preconcep-

tions and perhaps consider using three or even more cameras; the lesson for psychologists is that introspection is a very blunt scientific instrument.

It is not surprising, given the ease with which we see and the immense difficulty in emulating it, that rather little is known of our visual system and its actions. The root obstacle to progress has been the lack of a sufficiently general definition of what vision does. On the one hand it has been studied indirectly as the sense that is used when reading language. It has also been studied as the sense that can accomplish two-dimensional pattern recognition and judge whether patterns of markings are similar or different. It has been involved in studies of motor behaviour, concerned with the co-ordination and control of actions by visual information. In many psychophysical studies, it is used because it is there—the underlying philosophy being that the visual system exists to discriminate small differences between patterns of luminance.

A notable exception to these generalisations was the late David Marr, whose 1982 book has marked a new chapter in research. The central point that Marr has communicated so effectively is that vision is actually impossible unless one knows a great deal about how the environment is, or should be, constrained. Atoms tend to clump into relatively large, stable, smoothly surfaced bodies, for example. The general philosophy that I adopt in this essay is the same but with two differences. The first is a concern for detailed findings of quantitative psychophysical experiments as well as results from more cognitive approaches to vision. The second, which is important from the third chapter of this essay onwards, is a consideration of what is meant by the notions of visual measurements (e.g. of edge blur and position). These two differences lead to a very different appreciation of the initial stages of human vision from that proposed by Marr (e.g. 1976).

In the first chapter of this essay I start by defining the scope of the early levels of vision that are called Primal Sketch visual processing. Imagine a real optical image of the scene around you: This has variations in light intensity that arise because of the shape, colour, and texture of the surfaces in the scene, and the layout of the bodies in the scene, as well as the layout of the light sources. The Primal Sketch is concerned to recover and register as many of these variations in intensity as it can. Such a record of the image provides a rich source of information concerning the layout of bodies in the scene and their nature. The task of the Primal Sketch, therefore, is also to organise the information into a representation that is meaningful for further analysis and action. The first chapter of this essay examines the computational theory of how this may be done.

The second and third chapters of this essay explore the implications of the computational analysis for the interpretation of psychophysical data and models. Since this essay is principally an exposition of my own point of view, Chapter 2 is based on the MIRAGE model of Watt and Morgan (1985).

Chapter 3 addresses the question of just what it means to say that the visual system measures spatial aspects of the retinal image, such as the length and orientation of lines. The central concern in this chapter is the effects of distortions introduced by the processes of finding edges and lines, and how these distortions can be corrected or avoided. This is generally possible for image attributes that fall into reliable and characteristic distributions, for which error-correcting codes are applicable. Spatial position is an attribute for which this technique is not applicable. In Marr's scheme the position or visual direction of an edge is simply determined by the location of a zero-crossing in an image plane. This assumes good image geometry, an assumption that I question.

The fourth and fifth chapters then consider the issues of grouping and dynamic control, and I believe these to be the reward for the reader who has persevered so far. The line of argument up to this point stresses the computational requirements of the system and its psychophysical performance. The need for structured representation and dynamic control at a very early stage in the sequence of processes can be traced back logically to the nature of the physical environment. In the fifth chapter I observe a similarity between the behaviour of the Primal Sketch, in its new form, and the phenomena of visual attention.

The synthesis that results is recapitulated in the final chapter. I would claim that the strength of the link between the low-level approaches of psychophysics and computational theory and the high-level approaches of cognitive visual function lies in the logic of the arguments that indicate the computational need for control. It is a strong claim that the computational approach and the psychophysics have built a sufficiently constrained and powerful model of the early stages of vision to make it sensible and meaningful to enquire how much of the supposedly higher-level aspects of the psychology of vision can be accounted for thereby.

I think that it is now necessary to re-examine the usefulness of a distinction between high-level and low-level processes in vision. If high-level control of early processing stages is a feature of the system, to what extent can those early stages be regarded as low-level? The concept of high- and low-level aspects to a visual task is a rather different issue and is clearly a valid distinction. It does not follow that these will map onto distinct high- and low-level processing stages within the system.

Contents

Acknowledgements ix

Preface xi

1. **Introduction** 1
   Output Requirements of Visual Processing 2
   Constraints on the Input to the Primal Sketch 5
   Two Dimensions 16
   The Nature of the Representation 21
   Summary 25
   The First Problem 26

2. **A Model for the Primal Sketch** 27
   Spatial Filters 27
   The MIRAGE Algorithm 32
   Some of the Evidence concerning MIRAGE 42
   Summary 58
   The Second Problem 58

3. **Measurements, Metrics, and Distortions** 59
   Measuring Image Attributes 60
   Distortions in Edge Location 68
   The "Phenonemenal Phenomenon" and MIRAGE 73
   Summary 76
   The Third Problem 79

4. Calculating Values for Spatial Position with Grouping **81**
Errors that Propagate through an Absolute Representation 82
Grouping 90
MIRAGE and Grouping 96
A Dynamic MIRAGE 110
Summary 113
The Fourth Problem 114

5. Control of Primal Sketch Processing **115**
Grouping and Texture 116
The Time Sequence of Automatic Processing 119
Psychological Experiments on Automatic Processing 125
Control by Deliberate Intervention 129
Summary 136
The Fifth Problem 136

6. Synopsis: Low-level Vision as an Active Process **137**
The Three New Agreements 138
An Analogy for the Primal Sketch 140
Relationships between Low-level Vision and Cognition 140
Coda: The Logic of This Essay 143

References **146**

Indices **149**

1 Introduction

Light is the freely available messenger that allows us to sense remote objects in our environment without the need to interact with them directly. This is the modern view of vision that began with the Persian philosopher and physicist, Alhazen (or more properly, abu-'Ali Al Hasen ibn Al Haytham, 965–1039). Vision would have been easier to understand if the older view that light is emitted by the eye as a type of feeler, had been correct. Laser range finders work on this principle and are at present the only flawless way of measuring depth with light images. In the same way the colour of a surface, i.e. its reflectance, is easily computed if you know the position and nature of the source of light and the orientation of the surface.

The more difficult modern view, however, is the accepted version, and much of the rest of this essay will be concerned with the effects of unknown sources of light on an unknown arrangement of surfaces in the scene. Light is an uncertain messenger: It is not like a nice steady weight that can be reliably measured; it is a stream of random and largely independent particles, the photons, each of which has its own characteristic energy. The rate of arrival of photons at the eye is variable; we call this photon noise, and the variability depends on the mean rate. The mean rate of arrival is the intensity of the light. It is usually expressed as intensity per unit area, illuminance, which is similar to luminance, the intensity per unit area of emitted light. To avoid these cumbersome photometric units, I shall use the term grey-level to refer to the illuminance of the retina at an arbitrary small area.

In vision, measurements of the intensity of light sources are not of interest.

The source is incidental; the major interest is the disposition of reflecting surfaces in three-dimensional space. Light from the sources is reflected at surfaces, perhaps many times, and some eventually enters the eye. The grey-level or intensity at a particular place in the retinal image is determined by the output and positions of the sources, the reflectance, depth, and orientation of the surface imaged at that point with respect to the sources and the observer, and any mutual illumination of surfaces (i.e. light reflected from one surface to another, which is then illuminated directly from the source and also indirectly via the first surface).

If one knows all these details, then the grey-level at that particular point in the image can be calculated. The problem in vision is that these processes cannot be reversed: The grey-level on its own does not distinguish between the various factors causing it. In principle, any given retinal image could arise from an infinity of possible scenes, including a flat uniform surface illuminated by a patterned light source (the principle behind slide and cine projection). In practice, of course, we are very rarely faced with any operational ambiguity. The visual system generally manages to make a choice concerning the scene, and it is usually correct. This choice is made on the basis of assumptions concerning the most likely types of scenes. The scenes that we inhabit are generally constrained.

OUTPUT REQUIREMENTS OF VISUAL PROCESSING

Vision exists so that we can see what to do. Ultimately visual tasks require a full scene description in terms of the visible bodies, their shapes and sizes, their positions and motions, and their surface colours and markings. We are a long way from understanding how this is done, but we can, for simplicity's sake, break the process down into a number of sub-processes.

Marr's analysis of the architecture of low-level vision is currently the most widely used (see Marr, 1982), even though there are doubts about many details. Marr divided the process of vision down into three sub-processes, each of which delivers a representation for the next sub-processes. The three representations may be summarized as:

Primal Sketch:	A two-dimensional representation of significant grey-level changes in the image.
2.5D Sketch:	A partial three-dimensional representation recording surface distances from the observer.
Solid-model based representation:	A fully worked out volumetric representation of bodies in the scene.

There are significant concepts in this simple framework. A *representation* is a symbolic descriptor. It builds a description from a finite *alphabet of*

primitive symbols (such as "edge", "bar", "corner"), each having an associated attribute list (recording: size, orientation, contrast, for example) and a *grammar* or set of rules that will exactly and exclusively generate all valid *sentences* or scene descriptions. A *sketch* is the process that produces, analyses, and represents the data.

This essay is concerned only with the Primal Sketch, the reason being that there are several psychophysical and psychological studies that indicate that the Primal Sketch is far from dull and straightforward. Whereas Marr tended to regard it as an inflexible, automatic, memoryless process producing something rather like an edge map, I shall describe some evidence that points to high-level control and memory very early in the process. It will be argued that many of the visual attention phenomena have their roots, trunk, and some branches in the Primal Sketch, and that there is a particularly striking simplicity to the machinery that belies a wonderfully rich diversity of function.

Output Requirements of the Primal Sketch

Why have a Primal Sketch? Why not just have a 2.5D Sketch as the first stage? The motive for the existence of a separate Primal Sketch in Marr's work is relatively simple. The image itself has a great deal of information that is irrelevant to the 2.5D Sketch, and the Primal Sketch can be used to provide an economical representation. A second reason is that many of the computational problems in the 2.5D Sketch, such as stereomatching, would be hopelessly confounded by grey-levels rather than, for example, edge tokens. To these two reasons, one can add the obvious statement that many visual processes require a representation of the grey-level changes, not a 2.5D Sketch. Imagine how text would have to be written if we had no access to a Primal Sketch.

It seems reasonable to require that the Primal Sketch squeeze as much meaningful information out of the image as possible. Only part of this will be relevant for the 2.5D Sketch. One ultimate goal of vision is an understanding of the layout of bodies in three-dimensional space. The term *body* is used here to refer to a compact solid mass that remains coherent, at least over the time scale of perception. Our perceptual understanding of the scene is going to be in terms of objects, which do not necessarily correspond to bodies. A tree in winter is one body, but may be represented as a hierarchy of objects: its overall bulk, the trunk, and largest boughs, or individual twigs. The term *object* refers to a unit of perception. Bodies cause the input to vision; objects cause behaviour that is the output of vision.

The main task of the Primal Sketch is therefore to extract from the image all the relevant information about the layout and character of visible surfaces and to construct a convenient representation. Ideally the representa-

tion at this level of processing will be used in turn for all other subsequent processes, and so it must be a rich source of information. The Primal Sketch representation will be used to construct a depth representation, using for example information about occlusions, shading, texture gradients, and disparity differences (if there are two Primal Sketches, one per eye). The interesting parts of images concern the locations where surfaces become occluded, especially where surfaces occlude themselves by turning away from the observer, and the locations where surfaces bend even though they may remain visible, such as sharp creases.

Figure 1.1 shows a ground plan for a scene and marks the position of an observer, O. The scene has three walls, W_1–W_4, within which there are five bodies: an upright circular cylinder, C; two boxes with rectangular cross-sections, S, R; an upright block with triangular cross-section, T; and a sphere, B. Various lines-of-sight from the observer are also shown and each reaches a point of particular interest, which we might require the Primal Sketch to identify. For example, the cylinder occludes itself at points C_1 and C_2. There is no sudden change in the character of the cylinder surface here, but to the observer, these two points will appear to be distinctive as the edges of the cylinder because there is a discontinuity in surface depth from the observer. Very often such occluding edges also correspond to discontinuities in surface orientation, i.e. *creases* or *corners*, such as points R_1, R_3, and W_1. Creases and corners can also be imaged so that they do not correspond to the occluding edges of objects, as at points T_2 and R_2.

The top of Fig. 1.2 shows the equivalent range or depth map for the observer at point O. This might be a useful precursor to a full reconstruction of Fig. 1.1 because it records the distance from O to the nearest reflecting surface in each direction. Figure 1.2 also shows the variation in the orientation of the visible surfaces in the scene with respect to the observer at O. Surface orientation is important because it determines the surface luminance: Surfaces that are head-on to the source of light have a higher illumination level per unit surface area than those oblique to the source. Notice that the lines-of-sight from the observer to occluding edges correspond to sudden changes in range from the observer and sometimes also in orientation. The lines-of-sight to creases, such as T_2 and R_2, correspond to abrupt changes in surface orientation, but not in range.

We could start by requiring that the Primal Sketch discover as many points in the image that correspond to these interesting lines-of-sight as possible. We can go further and, expecting that many surfaces will have textured markings on them, ask that some measure of the variations in texture grain size over each surface patch be made, so that the shape and orientation of the surface and its edges can be subsequently assessed. These requirements can be summarised by stating that the Primal Sketch should produce a representation of the image in which features corresponding to surface creases and discontinuities are recorded, along with their shape,

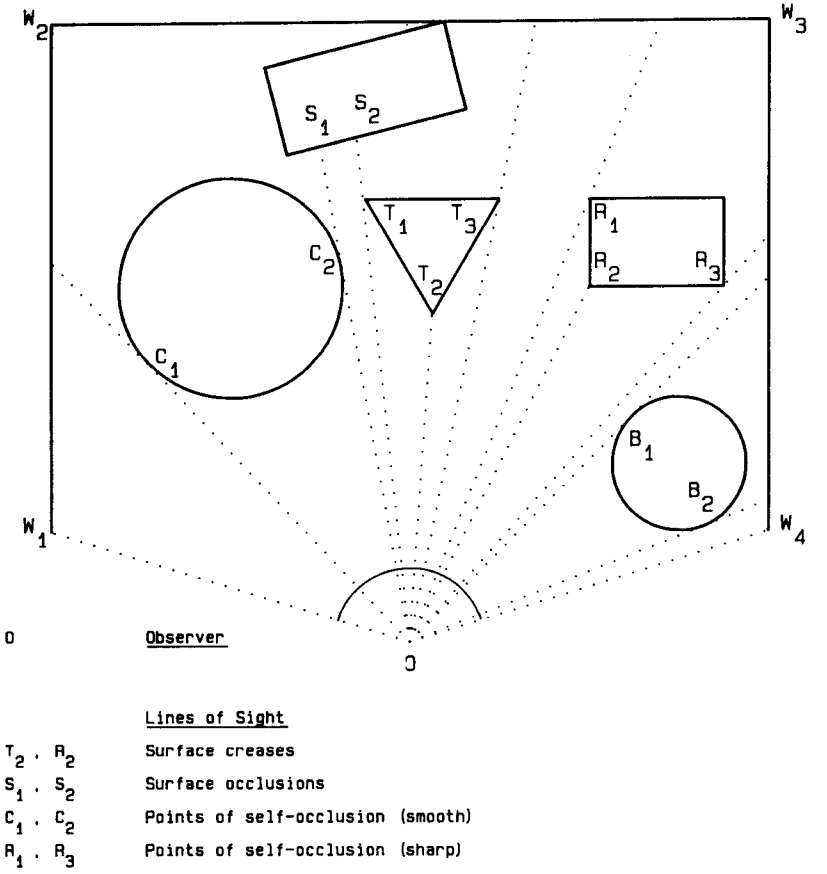

FIG. 1.1. A plan of an imaginary scene. The image formed by the observer at O will contain segments corresponding in sequence to the various visible surfaces. These segments are bounded by the lines-of-sight that are drawn from O to each point of surface occlusion or surface creasing. The problem for the Primal Sketch is to discover these lines-of-sight from the image and then to represent their spatial relations and their nature or probable cause (i.e. occlusion or crease).

orientation, and relative locations. The representation should also record variations in surface texture.

CONSTRAINTS ON THE INPUT TO THE PRIMAL SKETCH

The task just defined for the Primal Sketch would be hopelessly difficult except for the fact that natural surfaces are generally homogeneous in colour

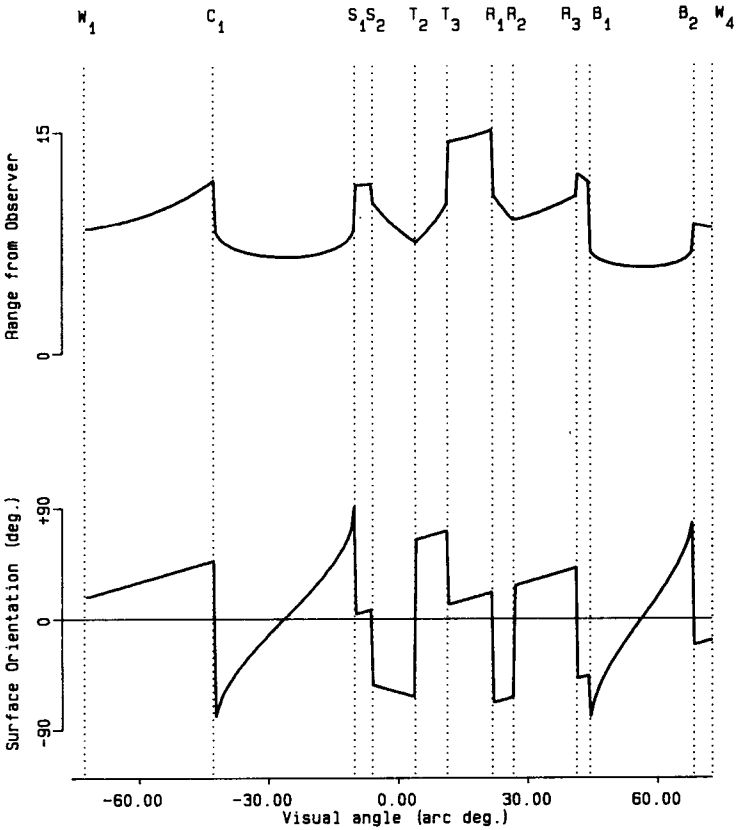

FIG. 1.2. A range map (top) and surface orientation map (bottom) for the scene of Fig. 1.1 from the point *O*. The different surface segments shown in that figure now correspond to areas where range and surface orientation change smoothly, and the segment boundaries are marked by sharp changes of direction and discontinuities in these maps.

and texture, and that these properties vary from surface to surface. As a result, any place in the image where the *nature of the reflected light changes suddenly* is likely to correspond to an occluding edge, although there are several caveats.

It is only usually the case that changes in reflected light indicate surface discontinuities or creases: Shadows very often cause a larger change in luminance than surface discontinuities, for example. However, the nature of the reflected light may not change very much at a shadow in terms of its spectral content or texture, which are predominantly determined by the surface itself. If the change in reflected light is sudden, then it is unlikely to be a shadow or a slowly varying gradient of illumination. But recall that all light

is carried by random photons. Adding photon noise to an image is the same as adding a random number to the grey-level value of every point in the image. There will always be sudden changes between adjacent points, so it is not sufficient to detect sudden changes in grey-level in the image. A phrase like "sudden and consistent on average" has to be employed instead to imply that surfaces generally occupy large areas of the image, whereas noise fluctuations are very restricted in size.

Luminance and Surfaces

Return to the scene of Fig. 1.1. How does the luminance of the various surfaces in the scene vary? Luminance depends on both the surface reflectance (colour), which is a property of the surface, and how it is illuminated. Let us suppose that they all have the same Lambertian reflectance, which is the most difficult condition for a visual system to deal with. All the surfaces are the same shade of matt grey. Let us also start with the case where there is a single point source of light at the position O. In this case there are no shadows visible to the observer.

Figure 1.3 shows the variations in grey-level along a one-dimensional slice through the image formed of the scene by the observer at O. These variations in grey-level are equivalent to variations in surface luminance, and in keeping with common practice, I shall use the terms luminance and grey-level synonymously when referring to an image. The variations in luminance arise because the distance of the surfaces from the source varies and because luminance is reduced by an amount depending on the angle of incidence of the light on the surface.

Inspection of Fig. 1.3 suggests that where surface creases and discontinuities exist, a sharp corner in the luminance profile can be expected. At such sharp corners, luminance slope changes suddenly (discontinuously). The first derivative of a function specifies its rate of change or slope at each point. It is found by taking each luminance value and subtracting the luminance value to be found a very small distance to its left (leftwards is only a convention; rightwards would produce the same result, multiplied by -1 and shifted very slightly). Discontinuities in the first derivative are places where there is a sudden large change in its value, and these can be easily detected by looking for isolated large deviations from zero (positive and negative) in the derivative of the first derivative, that is, in the second derivative of the luminance profile. The second panel of Fig. 1.3 shows the second derivative: For each of the marked scene features, there is an isolated peak and/or trough in this function.

To summarise, changes in luminance are important because they are related to surface orientation and reflectance. Where the rate of change itself changes suddenly, the visual system can learn about changes in surface

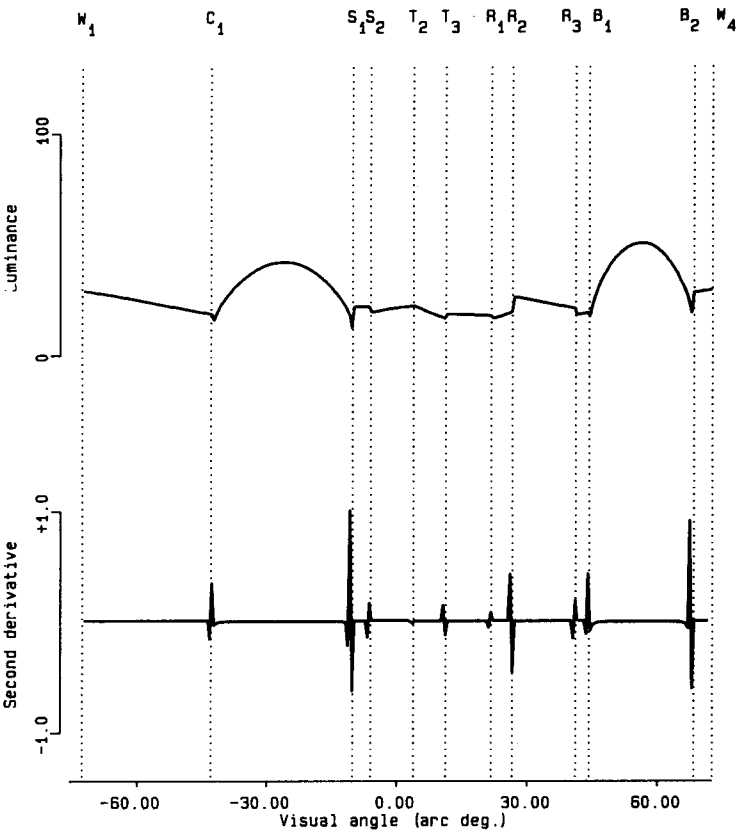

FIG. 1.3. The top of this figure shows the luminance profile of the image formed of the scene at the point *O*. The different surface segments shown in the figure now correspond to areas where luminance changes smoothly. The bottom of the figure shows the second derivative of the luminance profile. Surface creases and discontinuities are now marked by sharp peaks in this new signal.

orientation. Changes in a rate of change are measured by taking second derivatives.

This has been a consideration of a special case where the only source of light is illumination from a *point source*. How about more general conditions of illumination? For most scenes, there are a small number of relatively remote sources of light. Initially these sources provide *parallel illumination* as the light strikes the scene surfaces, so the surface luminances due to this depend on the direction of the light source. Light reflected off all the surfaces is scattered through the scene and provides secondary illumination of each surface. This is much less directional and effectively becomes *diffuse illumination*, which can be regarded as normal (i.e. perpendicular) to any and all