Technical Report 232

Shape From Shading: A Method for Obtaining the Shape of a Smooth Opaque Object From One View

Berthold K. P. Horn

MIT Artificial Intelligence Laboratory

SHAPE FROM SHADING: A METHOD FOR OBTAINING THE SHAPE OF A SMOOTH OPAQUE OBJECT FROM ONE VIEW

Berthold K. P. Horn

November 1970

PROJECT MAC

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Cambridge Massachusetts 02139

ACKNOWLEDGMENTS

The author wishes to express his appreciation to everyone who contributed to the interesting research environment.

Work reported herein was supported in part by Project MAC, an M.I.T. research program sponsored by the Advanced Research Projects Agency, Department of Defense, under Office of Naval Research Contract Number Nonr-4102(02). Reproduction in whole or in part is permitted for any purpose of the United States Government.

This research was also supported by the South-African Chamber of Mines and the Council for Industrial and Scientific Research.

SHAPE FROM SHADING: A METHOD FOR OBTAINING THE SHAPE OF A SMOOTH OPAQUE OBJECT FROM ONE VIEW*

Abstract

A method will be described for finding the shape of a smooth opaque object from a monocular image, given a knowledge of the surface photometry, the position of the light-source and certain auxiliary information to resolve ambiguities. This method is complementary to the use of stereoscopy which relies on matching up sharp detail and will fail on smooth objects. Until now the image processing of single views has been restricted to objects which can meaningfully be considered two-dimensional or bounded by plane surfaces.

It is possible to derive a first-order non-linear partial differential equation in two unknowns relating the intensity at the image points to the shape of the object. This equation can be solved by means of an equivalent set of five ordinary differential equations. A curve traced out by solving this set of equations for one set of starting values is called a characteristic strip. Starting one of these strips from each point on some initial curve will produce the whole solution surface. The initial curves can usually be constructed around so-called singular points.

A number of applications of this method will be discussed including one to lunar topography and one to the scanning electron microscope. In both of these cases great simplifications occur in the equations. A note on polyhedra follows and a quantitative theory of facial make-up is touched upon.

An implementation of some of these ideas on the PDP-6 computer with its attached image-dissector camera at the Artificial Intelligence Laboratory will be described, and also a nose-recognition program.

*This report reproduces a thesis of the same title submitted to the Department of Electrical Engineering, Massachusetts Institute of Technology, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, June 1970.

0. CONTENTS

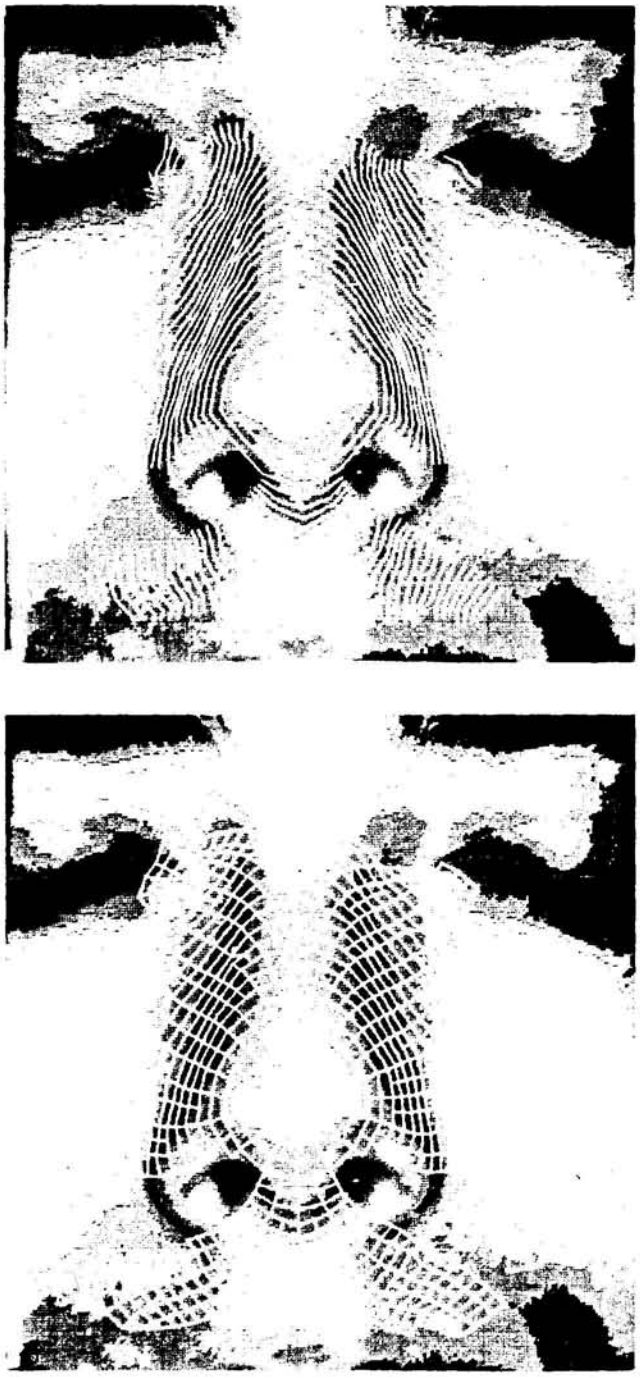


Figure 1: Pictures of a nose with superimposed characteristic solutions and contours. Shape determined from the shading (not intensity contours). See section 4.3 for details.

1. INTRODUCTION:

1.1 SHADING AS A MONOCULAR DEPTH CUE:

Consider a smooth object known to have a uniform surface. An image of such an object will exhibit shading (gradations of reflected light intensity) which can be used to determine its shape, given only a picture from a single viewpoint. This is not obvious since at each point in the image we know only the reflectivity at the corresponding object point. For some points (called singular points here) the reflectivity does uniquely determine the local normal, but for almost all points it does not. The shape of the surface cannot be found by local operations alone.

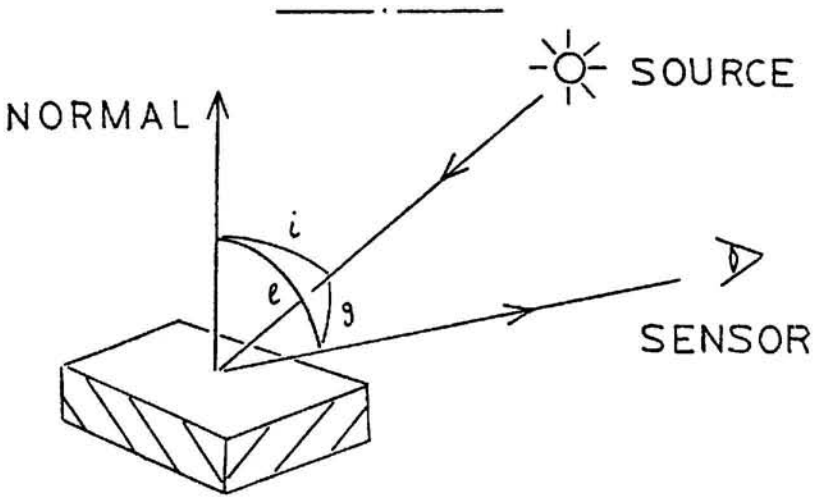


Figure 2: Definition of the incident (i), emittance (e) and phase angle (g).

For many surfaces the fraction of the incident light which is scattered in a given direction is a smooth function of the angles involved. It is convenient to think of the situation as depending on three angles: the incident angle (between local normal and incident ray), the emittance (or emergent) angle (between local normal and emitted ray) and the phase angle (between incident and emitted rays).

It can be shown that the shape can be obtained from the shading if we know the reflectivity function and the position of the light-source(s). The reflectivity and the gradient of the surface can be related by a non-linear first-order partial differential equation in two unknowns. The recipe for solving this equation is to set up an equivalent set of five ordinary differential equations (three for the coordinates and two for the components of the gradient) and then to integrate these numerically along certain curved paths on the object called characteristics [5]. For while we cannot determine the gradient locally, we can, roughly speaking, determine its component in one special direction. Then taking a small step in this direction, we can repeat the process - the curve traced out on the object in this manner is called a characteristic. Its projection on the image plane will be referred to as the base characteristic. The shape of the visible surface of the object is thus given as a

sequence of coordinates on some such curves along its surface.

An initial known curve on the object is needed to start the solution. Such a curve can usually be constructed near the singular points mentioned earlier using the known local normal. The only additional information needed is the distance to the singular point and whether the surface is convex or concave w.r.t. the observer at this point - such ambiguities arise in several other instances in the process of solution as will be seen.

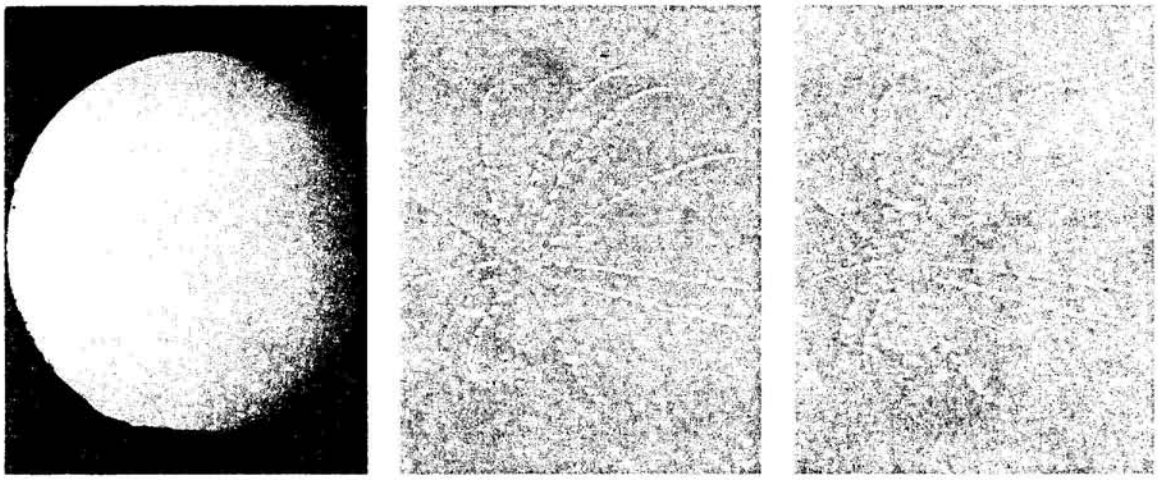


Figure 3: Image of a sphere and a stereo-pair of the characteristic curves obtained from the shading.

To solve the equations, the reflectivity as a function of the three angles must be known, as well as the geometry relating light-source, object and observer. Multiple or extended

light-sources increase the complexity of the solution algorithm presented. But all of this initially needed information can be deduced from the image if a calibration object of known shape is present in the same image. Furthermore, incorrect assumptions about the reflectivity function and the position of the light-source(s) can lead to inconsistencies in the solution and it may be possible to utilize this information in the absence of a calibration object.

In practice it is found that if the object is at all complex, its image will be segmented by edges. Some of these are purely visual, due to the occlusion of one surface by another, others are angular edges (also called joints here) on a single object. Another kind of edge is the ambiguity edge. This is an edge which the characteristics cannot cross, indicating an ambiguity which cannot be resolved locally. One can solve inside each region bounded by these various edges, but some global or external knowledge is needed to match up the regions. In the case of an angular edge on the object one can integrate up to the edge and then use the known location of the edge as an initial curve for another region (provided one resolves the ambiguity present here, as on all initial curves).

A very similar situation also obtains when one bridges a shadow. Since one edge of the shadow and the position of the light source is known, we can trace along the rays grazing the edge until the corresponding image points fall on an illuminated region. Since we know the path of each ray, we can calculate the coordinates of the point where it impinges on the object. The edge of the shadow (which need not be on the same object) can now serve as an initial curve from which to continue the solution.

A number of interesting applications of this method can be mentioned. The first of these concerns the scanning electron microscope (SEM) which produces images which are particularly easy to interpret, since the intensity recorded is a function of the slope of the object at that point and is thus a form of shading (as opposed to optical and transmission electron microscopes which produce intensities which depend on thickness and optical or electron density). The geometry of the scanning electron microscope allows several simplifications in the algorithm for determining shape from shading (e.g. there are no shadows). Because of the random access capability of the beam of this microscope it should be easy and useful to combine it with a small computer to obtain three-dimensional information directly.

Another important application lies in the determination of lunar topography. Here the special reflectivity function of the material in the maria of the moon allows a very great simplification of the equations used in the shape-from-shading algorithm. The equations in fact reduce to one integral which has to be evaluated along each of a family of predetermined straight lines in the image, making for high accuracy. This problem was first tackled for areas near the terminator (the dividing line between the illuminated and the unilluminated part of the moon's disk) by J. van Diggelen at the Astronomical Institute of the Netherlands in 1951 [2] and solved by T. Rindfleisch at the Jet Propulsion Laboratory in 1966 [4] and the method applied to several pictures returned by the Ranger spacecraft. This gave the first indication that the general solution discussed here might be possible.

It should be pointed out that this method is complementary to the use of stereopsis, since the latter will match up sharp detail and edges while the shading information will determine the shape of the smooth portions of the surface.

So far we have assumed that the surface is uniform in its photometric properties. Any non-uniformity will cause this algorithm to determine an incorrect shape. This is one of the uses of facial make-up; by darkening certain slopes they

can be made to appear steeper for example. In some cases surface-markings can be detected if they lead to discontinuities of the calculated shape.

Judging by our wide use of monocular pictures (photographs or even paintings and woodcuts) of people and other smooth objects, humans are good at interpreting shading information. The short-comings of our method which are related to the shading information available can be expected to be found in human visual perception too. It will of course be difficult to decide whether the visual system actually determines the shape quantitatively or whether it uses the shading information in a very qualitative way only. A quantitative determination would involve operations more complicated than those used in edge-finding for example. Since the information is not local, the surface-shape calculations cannot be carried out entirely in parallel.

1.2 HISTORY OF THE PROBLEM:

After formulating the image illumination equation as the basis of a method of finding shape from shading, a literature search was performed to see if a solution had previously been obtained. The literature on perception has only a few

conjectures on the possibility of determining shape from the monocular depth-cue of shading. Photogrammetry does not pay much attention to the reflectivity function, but only various integrals of it, measured by such devices as the integrating photometer. With few exceptions machine perception so far has been restricted to objects which can usefully be considered two-dimensional and objects bounded by planes (polyhedra).

The one relevant research was found in the paper on lunar topography by T. Rindfleisch [4] which gives complete details of a solution obtained in the form of an integral in the special case of the reflectivity function of the moon. This raised the hope that a general solution existed. The (x',y',r) coordinate system used in [4] leads to intractable equations - but we found a solution using a different coordinate system, (x',y',z). As a check the solution for lunar topography was rederived from this set of equations (Rindfleisch found his solution in quite a different manner - searching for predetermined curves in the image along which the surface can be found as some integral involving the measured image illumination). A first program (old SHADE) was then written which solved along one characteristic at a time using various predictor-corrector-modifier methods [7].

Another program (REFLEC) was used to measure the reflectivity function from a calibration sphere. Various short-comings of our image-dissector sensing device were affecting the accuracy of these measurements. Since very little was known about the characteristics of this device on other than theoretical grounds [9], a program (TEXTUR) was developed to measure various properties such as resolution, signal to noise ratio, drift, settling time, scatter and pinholes in the photocathode. An attempt was then made to provide software to compensate for some defects such as distortion and non-uniform sensitivity, using measurements from test patterns (DISTOR).

These techniques allowed an estimation of what accuracy can be achieved under optimal conditions. The program had numerous problems when dealing with objects other than simple convex ones (mostly because it solved each characteristic separately) and as must be apparent, was sensitive to imperfections in the sensing device (partly because of the way it obtained intensity gradients).

After the defects in the first program had been found, and a decision made to rewrite it, a great simplification of the main equations was found using a different coordinate system (x,y,z) and a slight extension of standard vector notation

(the voluminous equations for the inconvenient coordinate system $(x^{\bullet}, y^{\bullet}, z)$ are not reproduced here). An unfortunate but unimportant side-effect is an increase in the complexity of the derivation of the lunar topography integral. The new equations and numerous changes in the method of solution were incorporated in a new program (new SHADE) which was less sensitive to the various shortcomings of our image-dissector. This program can handle objects somewhat more complicated than its predecessor and solves all characteristics at the same time.

In parallel with the programming work, theoretical efforts were made to define and get around some of the difficulties of the method of shape from shading. Of particular interest were applications where the equations simplify greatly. Unfortunately the massive simplification found in the case of lunar topography is unique. Of most interest are the cases where we have some advance knowledge of the characteristics (for lunar topography they are completely independent of the image - for the scanning electron microscope they are paths of steepest descent).

1.3 PREVIEW OF CHAPTERS - GUIDE TO THE HURRIED READER:

References to articles and books listed at the end will be by numbers enclosed in brackets. Numbers contained in parentheses refer to sections and subsections in this work. In an attempt to be complete, a few subsections were included which will have only limited appeal to some readers; hence this guide.

Chapter 1 provides an introduction to the depth-cue of shading, its use in determining shape and its history. Chapter 2 develops the necessary equations in detail, starting with the definition of the reflectivity function. Subsections 2.1.2 to 2.1.4 and 2.2 can well be skipped by the hurried reader. In section 2.3 the partial differential equation is obtained, the vector differentiation notation introduced and an equivalent set of five ordinary differential equations derived. Section 2.3 is perhaps the most important section. Sections 2.12 to 2.16 deal with some miscellaneous implications and may be omitted without loss of continuity.

Chapter 3 describes in detail some practical situations where the special conditions encountered make use of the method of determining shape-from-shading particularly attractive.

Section 3.1 deals with the scanning electron microscope. The reader should be warned about the tedious derivation of the simple integral for the case of lunar topography in 3.2 . Omitting subsection 3.2.3 will avoid the bulk of the algebraic detail, and most of the conclusions will be found anyway in subsection 3.2.4 .

Chapter 4 describes the experiments carried out with the two programs (using the results developed in chapter 2) to obtain shapes from images projected on an image dissector camera attached to the PDP-6 computer in the Artificial Intelligence Laboratory. Section 4.1 deals with the less successful first program, and contains details on auxiliary routines. Section 4.2 deals with the second program which solves the characteristics in parallel and also uses the important sharpening process. Sections 4.1 and 4.2 are next in importance to section 2.3 .

Section 4.3 describes an application to a recognition task - that of nose-recognition. Section 4.4 contains an overall summary and conclusions about the capabilities of the method of shape-from-shading, with subsection 4.4.1 giving suggestions for future investigations. This is followed by a list of references.

2. THEORETICAL RESULTS:

2.1 THE REFLECTIVITY FUNCTION:

2.1.1 DEFINITION OF THE REFLECTIVITY FUNCTION:

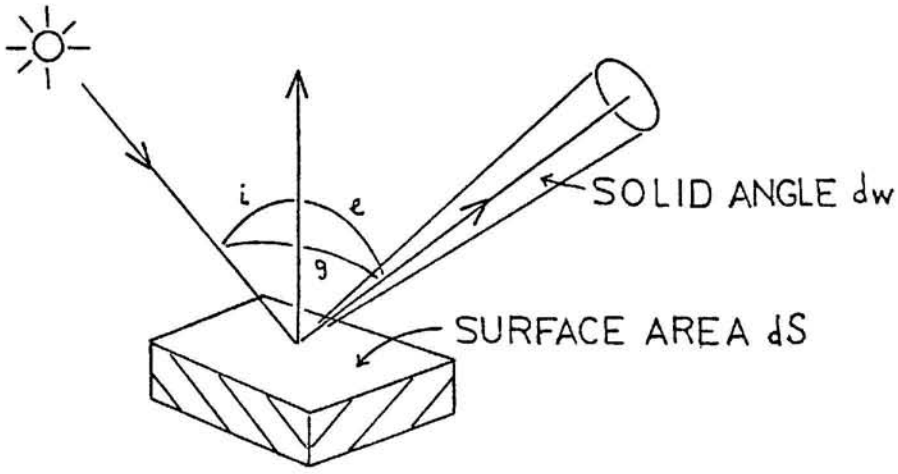
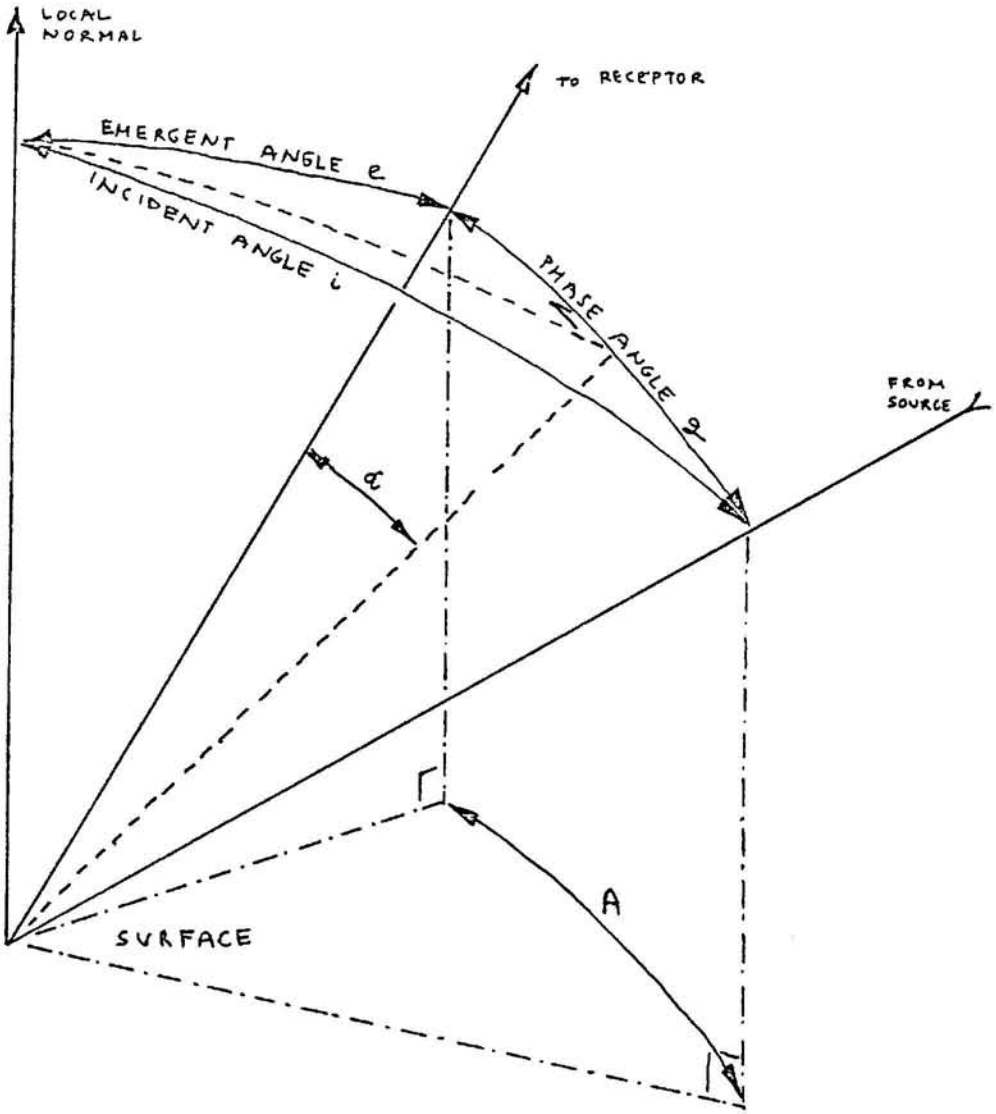


Figure 4: Illustration of the variables used in the definition of the reflectivity function.

Consider a surface element of size dS inclined i w.r.t. the incident ray and e w.r.t. the emitted ray (The angles are measured w.r.t. the normal). Let the incident light intensity be I_1 per unit area perpendicular to the incident ray. The amount of light falling on the surface element is then $I_1 \cos(i)\, dS$.

Let the emitted ray have intensity I_2 per unit solid angle per unit area perpendicular to the emitted ray. So the



$$\tan(\alpha) = \frac{\cos(e)\cos(g) - \cos(i)}{\cos(e)\sin(g)}$$

$$\cos(A) = \frac{\cos(g) - \cos(i)\cos(e)}{\sin(i)\sin(e)}$$

Figure 5: Definition of the azimuth angle (A) and the projection of the emittance angle on the phase-angle plane -(α).

amount of light intercepted by an area subtending a solid angle dw at the surface element will be $I_2 \cos(e)\, dS\, dw$. The reflectivity funcion $\phi(i,e,g)$ is then defined to be I_2/I_1.

If we want to be more precise about what units the intensity is measured in, we have to take into account the spectral distribution of the light emitted by the source, as well as the spectral sensitivity of the sensor (with this proviso we can speak of watts per unit area and watts per unit solid angle per unit area etc.). We need not be too concerned with this if we either use white paint, or measure the reflectiivity function with the same equipment later used in the shape-from-shading algorithm. It should be noted that for most surfaces the reflectivity function is not independent of the color of the light used. Typically the specular component of the reflected light, being reflected before it has penetrated far into the surface, will be unchanged, while the matt component will be colored by pigments in the surface coating.

Several other definitions of the reflectivity function are in use which are multiples of the one defined here by π, 2, cos(e) and/or cos(i). The specific formulation chosen here makes the equation relating the incident light intensity to the image illumination very simple.

2.1.2 FUNCTIONS DERIVED FROM THE REFLECTIVITY FUNCTION:

The next few subsections (2.1.2.1, 2.1.2.2 and 2.1.2.3) are included to relate the reflectivity functions to those more commonly mentioned in the literature. Some readers may want to skip these subsections.

2.1.2.1 THE INTEGRATING PHOTOMETER:

A flat sample of the surface under investigation is mounted in the center of a hollow sphere coated on the inside with a highly reflective matt substance. Through one small hole a light ray enters and impinges on the sample with incident angle i. A photosensitive device is introduced through another small hole and measures an intensity proportional to the light scattered by the sample into all directions.

The total intensity measured is:

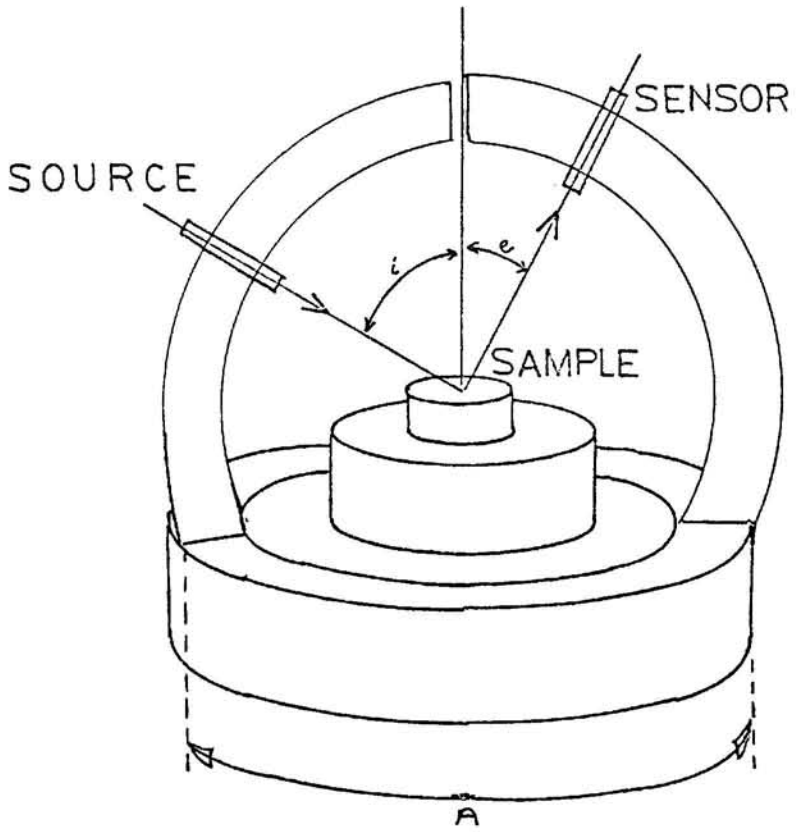


Figure 6: A gonio-photometer (used for measuring reflectivity functions).

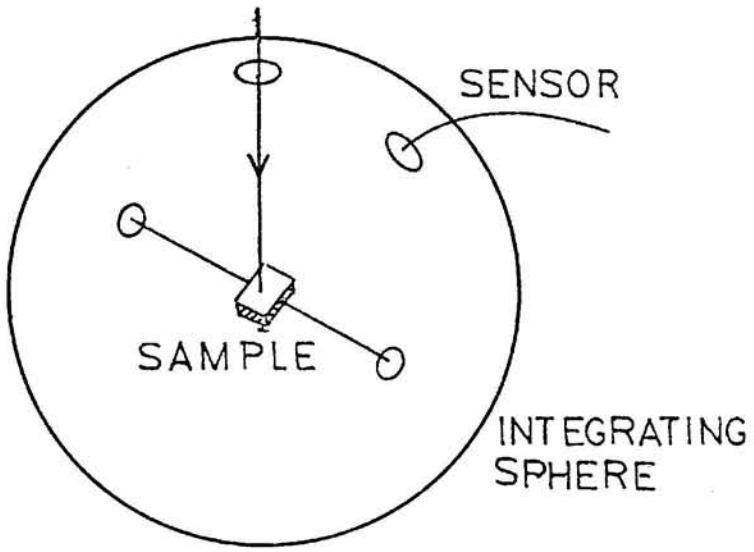


Figure 7: An integrating photometer.

$$\int_0^{2\pi}\int_0^{\pi/2} I_2 \cos(e)\ dS\ \sin(e)\ de\ dA$$

i.e.:
$$\int_0^{2\pi}\int_0^{\pi/2} \phi(i,e,g)\ I_1 \cos(e)\ dS\ \sin(e)\ de\ dA$$

Where $\cos(g) = \cos(i)\ \cos(e) + \sin(i)\ \sin(e)\ \cos(A)$

The total incident intensity is I cos(i) dS. The fraction of light reflected is then:

$$b(i) = \left[\int_0^{2\pi}\int_0^{\pi/2} \phi(i,e,g)*(1/2)*\sin(2e)\ de\ dA\ \right]/\cos(i)$$

This function of the incident angle i has been measured for many paints and pigments, while the reflectivity function ϕ is known for very few surfaces. Since it is difficult to relate measurements of I_1 to measurements of total reflected intensity, the device is usually calibrated with the sample replaced by a standard of known high reflectivity (e.g. MgO or $BaSO_4$ powder reflect more than 99% of the incident light in the visible spectrum).

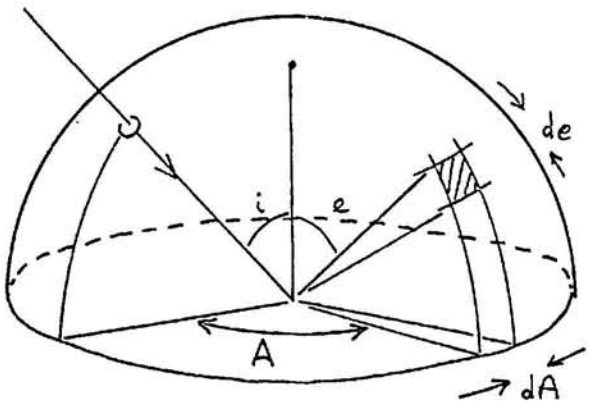


Figure 8: Illustration showing quantities appearing in the integral for the integrating photometer.

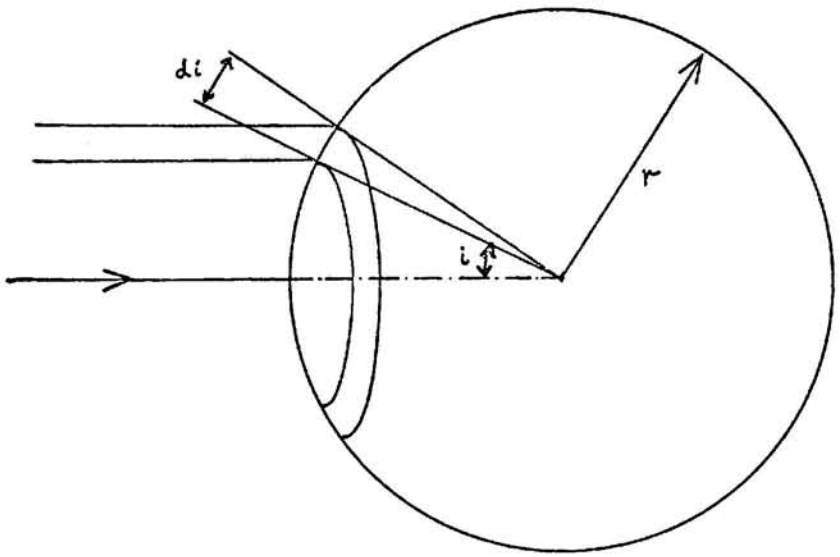


Figure 9: Illustration showing quantities appearing in the integral for the Bond albedo.

2.1.2.2 PERFECT DIFFUSERS - LAMBERT'S LAW:

Surfaces made of finely divided powder usually closely approximate what has been called a lambertian reflector or perfect diffuser. Lambertian emission was first defined for black-body radiation and is such that the surface of the body appears equally bright from all directions. In the context of reflectivity functions we call a surface lambertian if $\phi = k \cos(i)$ (the $\cos(i)$ accounts for the variation in incident radiation). For the highly reflective standards mentioned above, we chose k such that all the incident light is reflected.

$$k \int_0^{2\pi} \int_0^{\pi/2} (1/2) * \sin(2e)\ de\ dA = 1$$

$$k = 1/\pi$$

In addition to the various multiplicative factors shown above, a normalized reflectivity function is also used, where:

$$\rho \phi'(i,e,g) = \phi(i,e,g)$$

ρ is called the normal albedo and $\phi'(0,0,0) = 1$.

2.1.2.3 THE BOND ALBEDO:

Another integral of the reflectivity function which is used is the Bond albedo, defined by astronomers as the ratio of total reflected light from a sphere divided by the total incident light .

If the incident intensity is I_1, then the ratio of reflected light to incident light is:

$$b = [\, I_1 \int_0^{\pi/2} b(i)\; 2\pi r \sin(i) \cos(i)\; di \,]/(I \pi r^2)$$

$$= \int_0^{\pi/2} b(i) \sin(2i)\; di$$

$$= \int_0^{\pi/2} \int_0^{2\pi} \int_0^{\pi/2} \phi(i,e,g) \sin(i) \sin(2e)\; de\; dA\; di$$

2.1.3 THE DISCRIMINANT $1+2IEG-(I^2+E^2+G^2)$:

In this subsection a discriminant is developed which is needed in the program implementing the shape-from-shading algorithm. This section can be skipped without loss of continuity.

The three angles i, e and g, being the sides of a spherical triangle, have to satisfy the following relationships:

$$i+e \geqslant g,\ e+g \geqslant i \text{ and } g+i \geqslant e$$

It is often convenient to express these three relationships in terms of the cosines I, E and G of the three angles. We first note that only one of the relationships could fail at a time. For example if $i+e < g$:

$$i+2e < g+e \quad \text{i.e.} \quad i < g+e \text{ and}$$
$$2i+e < g+i \quad \text{i.e.} \quad e < g+i$$

The angles are all positive and less than π. Now assume that the condition $i+e < g$ holds, then:

$$\cos(i+e) > \cos(g)$$

since cosine is monotonic decreasing for angles between 0 and π. Expanding we get:

$$\cos(i)\cos(e) - \cos(g) > \sin(i)\sin(e)$$

Since the right-hand side is positive, the left-hand side will be too and we can square the expression. Using I, E and

G to stand for the cosines of i, e and g we get:

$$(IE-G)^2 > (1-I^2)(1-E^2) \text{ i.e.}$$
$$1+2IEG-(I^2+E^2+G^2) < 0$$

We now have to prove the converse i.e. if the angles can indeed form a spherical triangle then the discriminant will be positive. Since $i \leqslant e+g$ we have $g \geqslant i-e$ and similarly since $e \leqslant i+g$ we have $g \geqslant e-i$, so:

$$g \geqslant |e-i| \text{ and similarly } i \geqslant |g-e| \text{ and } e \geqslant |i-g|$$

Applying the cosine as before we get:

$$\cos(g) - \cos(i)\cos(e) \leqslant \sin(i)\sin(e)$$
$$\cos(i) - \cos(e)\cos(g) \leqslant \sin(e)\sin(g)$$
$$\cos(e) - \cos(g)\cos(i) \leqslant \sin(g)\sin(i)$$

From $i+e \leqslant g$ etc. we get the same inequalities with the sign reversed on the left-hand side. We needed to go to the trouble of showing that these inequalities hold for absolute values of the left-hand side, since it no longer is constraint to be positive. So we have:

$$|IE-G| \leqslant \sin(i)\ \sin(e)$$

$$|EG-I| \leqslant \sin(e)\ \sin(g)$$

$$|GI-E| \leqslant \sin(g)\ \sin(i)$$

Multiplying we get:

$$|(IE-G)(EG-I)(GI-E)| \leqslant (1-I^2)(1-E^2)(1-G^2)$$

Using one of the two signs for the right-hand side and expanding we get:

$$(1-IEG)(1+2IEG-(I^2+E^2+G^2)) \geqslant 0$$

Since $|I|$, $|E|$ and $|G| \leqslant 1$ we have $(1-IEG) \geqslant 0$ and hence:

$$1+2IEG-(I^2+E^2+G^2) \geqslant 0$$

2.1.4 REFLECTIVITY FUNCTIONS AND THEIR MEASUREMENT:

Surfaces where the three parameters i, e and g are not sufficient to fully determine the reflectivity are unsuitable for this analysis (or at least reduce the possible accuracy). Examples are translucent objects and those with non-isotropic surface properties (e.g. the mineral commonly called tiger-

eye, hair, thin wax). Perhaps the most important determinant of the reflectivity function is the micro-structure of the surface (i.e. that structure smaller than the resolution of the sensor used in the determination of the reflectivity) and different reflectivity functions may apply at different magnifications (in addition, at high magnification objects become increasingly translucent). It is best then to determine the reflectivity function under conditions similar to those used in the determination of the shape of the object.

One way to measure the reflectivity function is to employ a gonio-photometer fitted with a small flat sample of the surface to be investigated. The device can be set for any combination of incident, emittance and phase angles.

To avoid having to move the source and the sensor into all possible positions w.r.t. a flat sample of the surface when measuring the reflectivity function, it is convenient to have a test-object which presents all possible values of i and e for a given g. (The constraints are $i+e \leqslant g$, $e+g \leqslant i$ and $g+i \leqslant e$). Use of such an object is greatly simplified by using a telephoto lens and a distant source, giving almost constant g. It is convenient to tabulate the reflectivity versus i and e for each of a series of values of g. A sphere

is perhaps the easiest test-object to use if one is willing to live with the decreasing accuracy in determining e as one approaches the edge.

One could also have an object of known shape in the same scene as the object to be analysed. This solves the problem of having to know the source location and the transfer properties of the image forming system as well. In some cases objects of known shape and surface characteristics differing from those of the object under study are useful - for example a sphere with specular reflectivity can pinpoint the location of the light-sources (e.g. the eyes in a face).

2.1.5 MATHEMATICAL MODELS OF SURFACES:

A number of attempts have been made to predict reflectivity functions on a theorectical basis starting with some assumed micro-structure of the surface. White matt surfaces are usually finely divided grains of transparent material (e.g. snow and crushed glass). White paint usually consists of transparent 'pigment' particles (e.g. of SiO_2 or TiO_2) of high refractive index (1.7 to 2.8) and small size (optimally about the wavelength of visible light) suspended in a transparent medium of low refractive index (1.3 say). If one

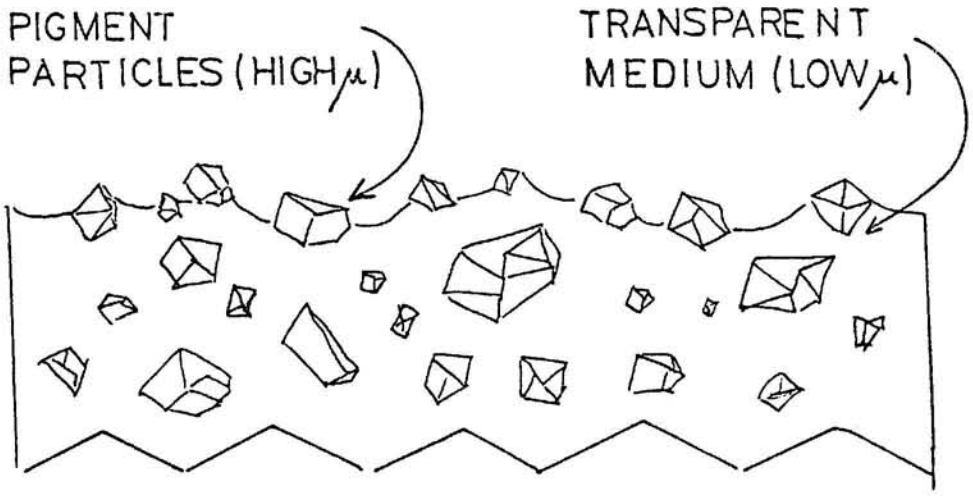


Figure 10: Model of surface-structure.

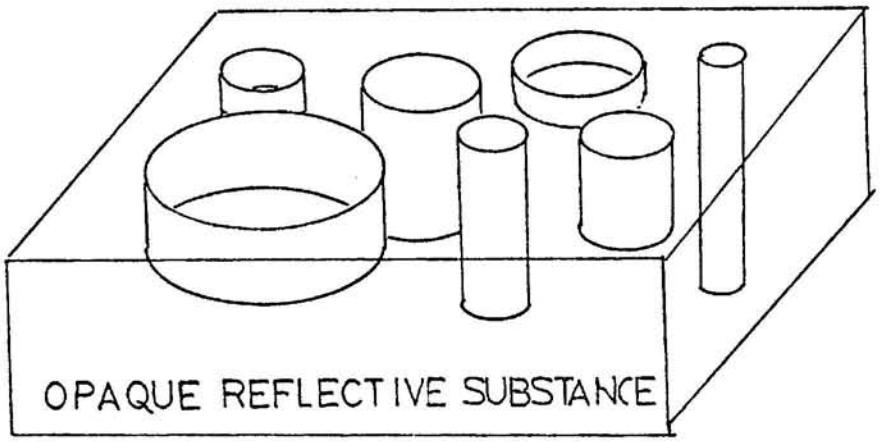


Figure 11: Another model used in the derivation of theoretical reflectivity functions.

choses a somewhat regular arrangement of suspended particles of uniform size and makes some very restrictive assumptions, one can derive a reflectivity function and study its dependence on various parameters describing the model of the surface.

Another type of surface is that of a highly reflective material (such as a metal) where the light rays do not penetrate into the material. Choosing a particular type of surface depression and a statistical distribution of the size of these, one can again derive a reflectivity function. Only a few such models have been studied and little hope exists for modelling real surfaces well enough and still deriving a closed expression for the reflectivity function.

2.2 CALCULATION OF IMAGE ILLUMINATION:

The equations derived in this section are only included for reference, since the program to be described later avoids their use by means of a normalization of the image intensity. These equations do have their importance however in justifying the choice of definition for the reflectivity function and in designing optical systems used in experimentation with the shape-from-shading method.

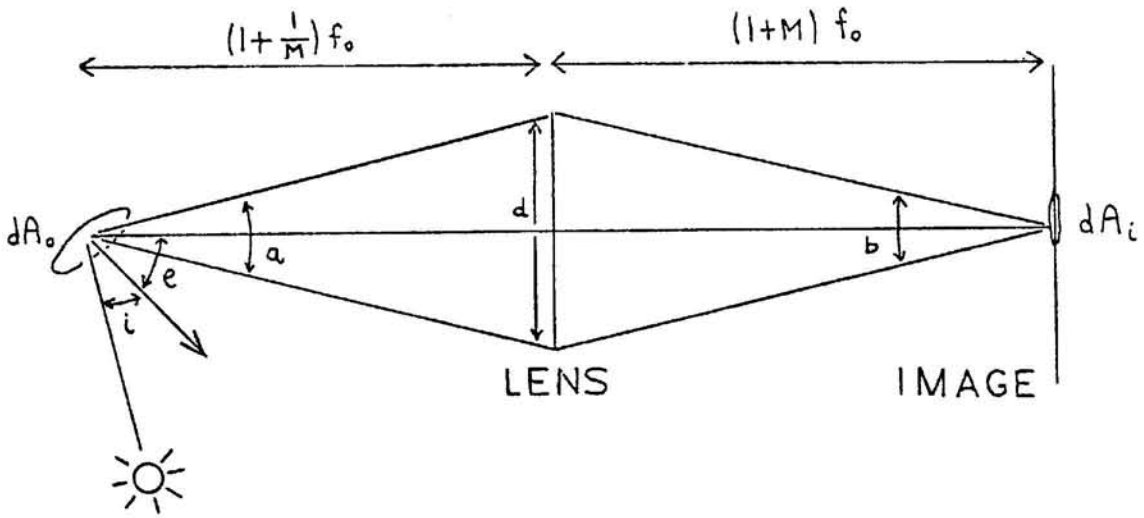


Figure 12: Diagram of optical system and quantities needed in the calculation of image illumination.

Let d be the diameter of the pupil of the image forming device, f_o its focal length and M the image magnification (the ratio of the length of a line in the image to the corresponding parallel line on the object).

Let a portion of the object surface of area dA_o be inclined at angle e to the line from it to the image-forming device. Its image will have an area of $dA_i = dA_o M^2/\cos(e)$.

Let the incident intensity at the object patch be I_i per unit area perpendicular to the incident ray. Then the emergent intensity per unit solid angle will be $I_2 = I_1\ \phi(i,e,g)$. The light captured by the image forming device is $I_2\ dA\ dw/\cos(e)$ where dw is the solid angle formed by the cone of angle a.

$$dw = 2\pi(1-\cos(a/2)) = 4\pi \sin(a/4)$$

We would like to express this in terms of M and the so-called f-number f_n:

$$f_n = 1/(2\sin(b/2))$$
$$= \sqrt{(1/4)+(1+M)^2(f_o/d)^2} \simeq (1+M)(f_o/d) \text{ if } (f_o/d)>1$$

The f-number usually indicated on the lens-barrel is (f_o/d) or $\sqrt{(1/4)+(f_o/d)^2}$.

$$f_n^2-(1/4) = (1+M)^2(f_o/d)^2$$

$$\cos(a/2) = \frac{(1+M)(f_o/d)}{\sqrt{(M^2/4)+(1+M)^2(f_o/d)^2}}$$

$$\cos(a/2) = \sqrt{\frac{(4f_n^2-1)}{(4f_n^2+M^2-1)}}$$

$$dw = 2\pi\left(1-\sqrt{1-\frac{M^2}{(4f_n^2+M^2-1)}}\right)$$

$$\simeq 2\pi\frac{M^2}{(4f_n^2+M^2-1)} \quad \text{if } f_n > 1$$

The intensity per unit area in the image is:

$$I_3 = I_2\,[dA_o/(\cos(e)\,M^2)]*[\cos(e)/dA\,]\,dw$$
$$= I_2\,dw/M^2 \simeq I_2\,2\pi/(4f_n^2) \text{ if } M < 1$$
$$I_3 = I_1\,\phi\,dw/M^2 \simeq I_1\,\phi\,2\pi/(4f_n^2) \text{ if } M < 1$$

It becomes apparent why we chose to define the reflectivity function the way we did and also why one might want factors of π and/or 2 in the definition. It should be noted that in practice one does not usually employ this equation, but rather normalizes the expressions used.

2.3 THE IMAGE ILLUMINATION EQUATION:

This section contains the derivation of the image illumination equation and the analytical formulation of the shape-from-shading problem.

2.3.1 PREVIEW OF HOW TO OBTAIN THE PARTIAL DIFFERENTIAL EQUATION:

At a known point on the object we can calculate g. We would like to find the gradient (or at least its component in one direction) at this point so as to be able to continue the solution to a neighboring point. Measurement of the light

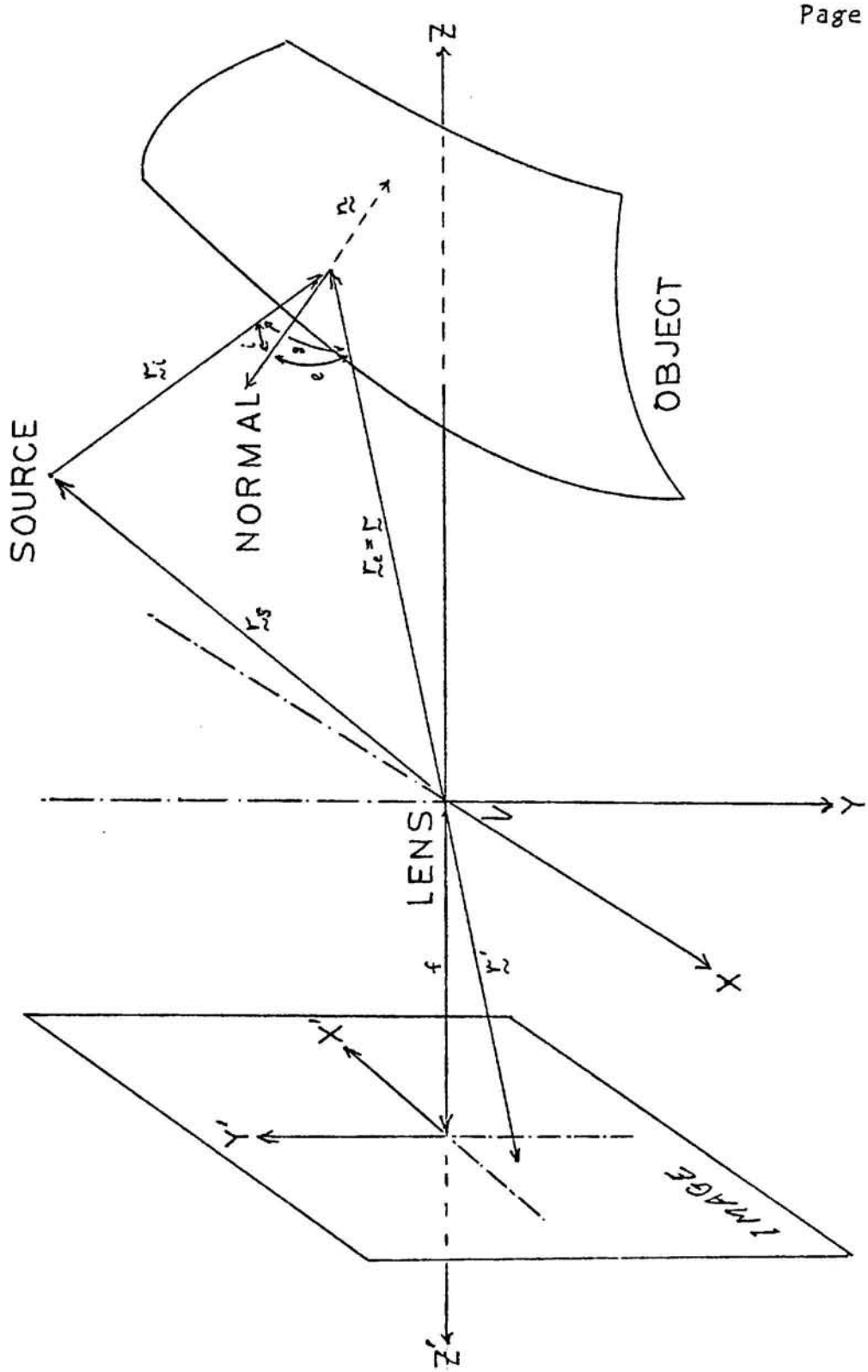


Figure 13: Details of the geometry of image illumination and projection in the imaging system.

reflected tells us something about i and e. Since only one measurement is involved, we cannot generally hope to determine both i and e locally, but only a relation between the two. There are exceptional points where the normal is locally fully determined and this is useful in finding initial conditions as explained later. This situation is contrary to that obtaining in the use of texture gradients (see section 2.16) where the gradient is known locally (except for a two-way ambiguity). In obtaining a solution from the shading, only a global approach will work.

Collapse the two principal planes of the image-forming system together, forming the x-y plane. Let the z-axis coincide with the optical axis and extend toward the object. Let f be the exit pupil to image plane distance and assume that the image and object space refractive indexes are equal.

Let t be the ratio of image illumination to object luminance (can be found from laws of optics - see section 2.2). Let a(x,y,z) be the incident light intensity (usually constant or obeys some inverse square law). Let A(x,y,z) = t*a(x,y,z)

Let $\underset{\sim}{r}$ = (x,y,z) be a point on the object and $\underset{\sim}{r}'$ = (x',y',f) the corresponding point in the image.

$b(x',y')$ is the intensity measured at the image point (x',y').

Let $I = \cos(i)$, $E = \cos(e)$ and $G = \cos(g)$.

We have $\quad A(\underset{\sim}{r})\ \phi(I,E,G) = b(\underset{\sim}{r}')$

Let p and q be the partial derivatives of z w.r.t. x and y. We would like to show that this equation involves x, y, z, p and q only.

2.3.2 NOTATION FOR VECTOR DIFFERENTIATION:

If $\underset{\sim}{A}$ is a vector (3-tuple), then $A = |\underset{\sim}{A}|$ is the magnitude of $\underset{\sim}{A}$. Also let $\hat{\underset{\sim}{A}} = \underset{\sim}{A}/A$ be the corresponding unit vector. Consider the dot-product $\underset{\sim}{A}.\underset{\sim}{B}$ as matrix multiplication of the 1 by 3 matrix $\underset{\sim}{A}$ by the 3 by 1 matrix $\underset{\sim}{B}^T$ (the transpose of B). Consider partial differentiation w.r.t. a vector as the 3-tuple whose components are found by differentiating w.r.t. each component in turn. Then for example:

$$A_{\underset{\sim}{A}} = \hat{\underset{\sim}{A}}$$

At times we will also need the partial derivatives of vectors w.r.t. vectors. These are defined as 3 by 3 matrices (the first row being the result of differentiating w.r.t. the first component and so on), then for example:

$$\underset{\sim}{A}_{\underset{\sim}{A}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

We will also use partial derivatives of dot-products of unit vectors w.r.t. vectors. For example:

$$X = \hat{\underset{\sim}{A}}.\hat{\underset{\sim}{B}} \qquad \text{and we want } X_{\underset{\sim}{A}}$$

To avoid finding $\hat{\underset{\sim}{A}}_{\underset{\sim}{A}}$ we write $A\,X = \underset{\sim}{A}.\hat{\underset{\sim}{B}}$ and then:

$$A_{\underset{\sim}{A}}X + A\,X_{\underset{\sim}{A}} = \underset{\sim}{A}_{\underset{\sim}{A}}.\hat{\underset{\sim}{B}}$$

Extending the definition of dot-product in the appropriate way we find:

$$\underset{\sim}{A}_{\underset{\sim}{A}}.\hat{\underset{\sim}{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \hat{\underset{\sim}{B}}^{T} = \hat{\underset{\sim}{B}}$$

$$A\,X_{\underset{\sim}{A}} = \hat{\underset{\sim}{B}} - X\,\hat{\underset{\sim}{A}}$$

$$X_{\underset{\sim}{A}} = (1/A)(\hat{\underset{\sim}{B}} - X\,\hat{\underset{\sim}{A}})$$

2.3.3 THE EQUATION IS A FIRST-ORDER NON-LINEAR P.D.E.:

If the reflectivity function is $\phi(I,E,G)$, the normalized incident light intensity at the point $\underset{\sim}{r} = (x,y,z)$ is $A(\underset{\sim}{r})$ and the intensity at the corresponding image point $\underset{\sim}{r}' = (x',y',f)$ is $b(\underset{\sim}{r}')$, then:

$$A(\underset{\sim}{r})\ \phi(I,E,G) = b(\underset{\sim}{r}')$$

This image illumination equation is the main equation studied here. When finding a solution we assume $A(\underset{\sim}{r})$ and $\phi(I,E,G)$ are known and $b(\underset{\sim}{r}')$ is obtained from the image. We want to show that the equation is a first-order non-linear partial differential equation in two independent variables, i.e. an equation of the form:

$$F(x,\ y,\ z,\ p,\ q) = 0$$

where $p = z_x$ and $q = z_y$ are the partial derivatives of z w.r.t. x and y respectively. From the simple projection geometry we have:

$$\underset{\sim}{r}' = (f/z)*\underset{\sim}{r}$$

Where f is the exit pupil to image plane distance. We took care of image reversal by orienting the x' and y' axes appropriately. It remains to show that I, E and G are functions of x, y, z, p and q. An inward normal to the surface at the point $\underset{\sim}{r}$ is $\underset{\sim}{n} = (-p, -q, 1)$.

Let the light-source be at $\underset{\sim}{r}_s = (x_s, y_s, z_s)$. Then the incident ray will be $\underset{\sim}{r}_i = \underset{\sim}{r} - \underset{\sim}{r}_s$, and the emergent ray $-\underset{\sim}{r}_e = -\underset{\sim}{r}$. Clearly then:

$$I = \hat{\underset{\sim}{n}} \cdot \hat{\underset{\sim}{r}}_i, \quad E = \hat{\underset{\sim}{n}} \cdot \hat{\underset{\sim}{r}}_e \quad \text{and} \quad G = \hat{\underset{\sim}{r}}_i \cdot \hat{\underset{\sim}{r}}_e$$

Where the $\hat{}$'s denote unit vectors. All the terms thus involve only x, y, z, p and q. It follows that we are dealing with a first-order non-linear partial differential equation in the two unknowns x and y.

2.3.4 SOME DERIVATIVES NEEDED IN THE SOLUTION:

When solving the P.D.E. by the method of characteristics we will need the following partial derivatives (see section 2.5), which it is convenient to introduce here, following the expansion of I, E and G in terms of dot-products. Using the results developed in subsection 2.3.2 we get:

$$\underset{\sim}{I}_{\underset{\sim}{r}} = \underset{\sim}{I}_{\underset{\sim}{r}_i} = (1/r_i)(\hat{\underset{\sim}{n}} - \underset{\sim}{I}\,\hat{\underset{\sim}{r}}_i)$$
$$\underset{\sim}{I}_{\underset{\sim}{n}} = (1/n)(\hat{\underset{\sim}{r}}_i - \underset{\sim}{I}\,\hat{\underset{\sim}{n}})$$

$$\underset{\sim}{E}_{\underset{\sim}{r}} = \underset{\sim}{E}_{\underset{\sim}{r}_e} = (1/r_e)(\hat{\underset{\sim}{n}} - \underset{\sim}{E}\,\hat{\underset{\sim}{r}}_e)$$
$$\underset{\sim}{E}_{\underset{\sim}{n}} = (1/n)(\hat{\underset{\sim}{r}}_e - \underset{\sim}{E}\,\hat{\underset{\sim}{n}})$$

$$\underset{\sim}{G}_{\underset{\sim}{r}} = \underset{\sim}{G}_{\underset{\sim}{r}_i} + \underset{\sim}{G}_{\underset{\sim}{r}_e} = (1/r_e)(\hat{\underset{\sim}{r}}_i - \underset{\sim}{G}\,\hat{\underset{\sim}{r}}_e) + (1/r_i)(\hat{\underset{\sim}{r}}_e - \underset{\sim}{G}\,\hat{\underset{\sim}{r}}_i)$$
$$\underset{\sim}{G}_{\underset{\sim}{n}} = 0$$

2.3.5 THE EQUIVALENT SET OF ORDINARY DIFFERENTIAL EQUATIONS:

The usual method of dealing with a first-order non-linear partial differential equation is to solve an equivalent set of five ordinary differential equations:

$$\dot{x} = F_p, \quad \dot{y} = F_q, \quad \dot{z} = pF_p + qF_q$$
$$\dot{p} = -F_x - pF_z \quad \text{and} \quad \dot{q} = -F_y - qF_z$$

The dot denotes differentiation w.r.t. s, a parameter which varies with distance along a characteristic strip. The subscripts denote partial derivatives. These equations are solved along so-called characteristic strips (see [5], page 24). The characteristic strip are the characteristic curves

described earlier (values of x, y and z) plus the values of p and q on them.

Since we can multiply the equation $F = 0$ by any non-zero smooth function $\lambda(x, y, z, p, q)$ without altering the solution surface, we can obtain a different set of equations:

$$\dot{x} = \lambda F_p, \quad \dot{y} = \lambda F_q, \quad \dot{z} = \lambda(pF_p + qF_q)$$
$$\dot{p} = \lambda(-F_x - pF_z) \quad \text{and} \quad \dot{q} = \lambda(-F_y - qF_z)$$

The solution to this new set of equations will differ only in the values of the parameter s at any given point. For example if we let

$$= 1/\sqrt{F_p^2 + F_q^2 + (pF_p + qF_q)^2}$$

the parameter s gives us arc-length along the characteristics. This is used in the programs to be described later. Of course we can only do this if the denominator is not zero; at singular points and ambiguity edges it will be zero (i.e. $F_p = F_q = 0$ and $\lambda \to \infty$). A different choice for λ will be used later in the discussion of the scanning electron microscope (section 3.1).

2.3.6 OUTLINE OF PROOF OF EQUIVALENCE OF THE SET OF O.D.E.'S TO THE P.D.E.:

In this subsection the equivalence of the five ordinary differential equations to the image illumination equation is discussed. The reader who believes the equivalence holds may well skip this subsection!

At a given point (x_o, y_o, z_o) the equation $F(x, y, z, p, q) = 0$ represents a relation between p and q. That is, it confines the possible solution normals at this point to a one-parameter family of directions [5]:

$$(-p(\alpha), q(\alpha), -1)$$

Increments in the feasible tangent planes thus satisfy:

$$dz = p_o(\alpha)\, dx + q_o(\alpha)\, dy$$

Differentiating w.r.t. α we get:

$$0 = p_o'(\alpha)\, dx + q_o'(\alpha)\, dy$$

(Dashes are derivatives w.r.t. α in this subsection). But by differentiating the equation $F(x, y, z, p, q) = 0$ w.r.t. α we also get:

$$F_p p_o'(\alpha) + F_q q_o'(\ ) = 0$$

Hence: $$dx/F_p = dy/F_q$$

What we need now are similar equations for feasible increments in z, p and q. First we have:

$$dz = p\,dx + q\,dy$$

in the solution surface (this surface is selected from all possible ones by choosing one passing through a given initial curve - see later, subsection 2.3.7). Hence:

$$dz/(pF_p + qF_q) = (p\,dx + q\,dy)/(pF_p + qF_q)$$
$$= dx/F_p = dy/F_q$$

Finally differentiating $F(x, y, z, p, q) = 0$ w.r.t. x and y:

$$F_x + F_z p + F_p p_x + F_q q_x = 0$$
$$F_y + F_z q + F_p p_y + F_q q_y = 0$$

but $p_y = q_x$ $$\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x}$$

$$dp/(F_x + pF_z) = -(p_x dx + p_y dy)/(p_x F_p + p_y F_q) = -dx/F_p$$
$$dq/(F_y + qF_z) = -(q_x dx + q_y dy)/(q_x F_p + q_y F_q) = -dy/F_q$$

$$dx/F_p = dy/F_q = dz/(pF_p + qF_q)$$
$$= -dp/(F_x + pF_z) = -dq/(F_y + qF_z)$$

Introducing the parameter s we get the five O.D.E.'s mentioned earlier. We have shown that a solution to the P.D.E. must also satisfy theses five O.D.E.'s. It is a bit more difficult to show that a solution to these O.D.E.'s is necessarily a solution of the P.D.E. (see [5], page 28). Basically it needs to be shown that the equations for p and q produce results which continue to be consistent (i.e. equal to the partial derivatives of z w.r.t. x and y).

2.3.7 INITIAL CONDITIONS NEEDED:

To select a particular solution surface amongst all possible solution surfaces one needs to specify an initial curve through which the solution surface must pass:

$$x = x(t),\ y = y(t) \text{ and } z = z(t)$$

Along this curve we must satisfy:

$$z'(t) = p\,x'(t) + q\,y'(t)$$

$$F(x(t),\ y(t),\ z(t),\ p(t),\ q(t)) = 0$$

Here the dash represents differentiation w.r.t. t. This pair of non-linear equations allows one to find p(t) and q(t) along the initial curve (there may be more than one solution, in which case there will be more than one solution surface). The characteristic strips sprout from this initial curve and the solution surface can be described parametrically:

$$x = x(s,t),\ y = y(s,t),\ z = z(s,t) \text{ and}$$

$$p = p(s,t),\ q = q(s,t)$$

2.4 SIMPLIFYING CONDITIONS AND UNIFORM ILLUMINATION:

Since the general equations are fairly complex it is of great interest to find simplifying conditions. Some of these are presented here, others will be found described in chapter 3.

1. DISTANT SOURCE: (Collimated source or the object subtends a small angle at the source)

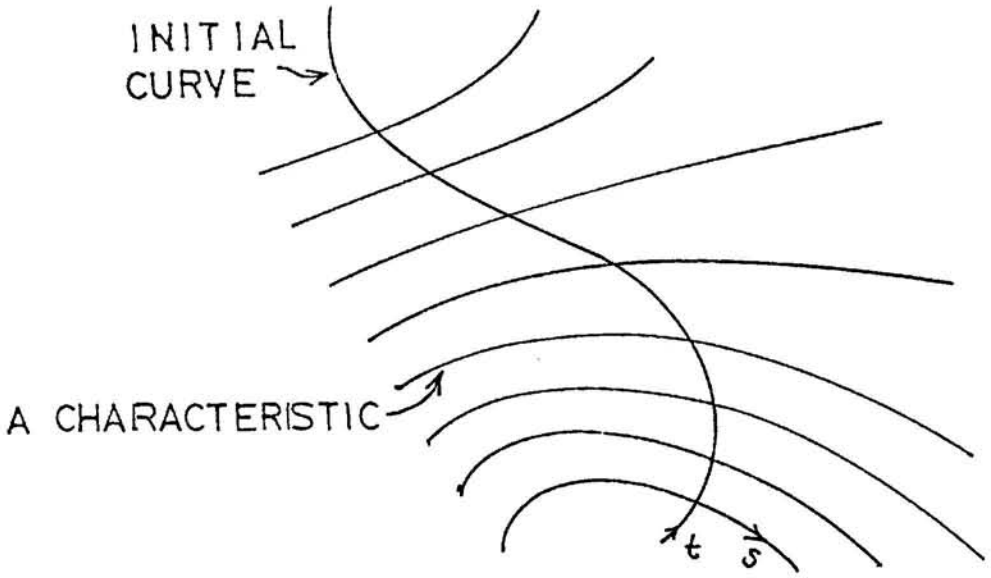


Figure 14: Characteristic strips sprouting from an initial curve.

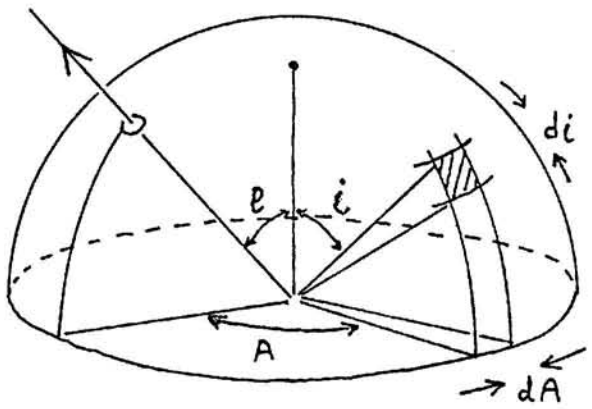


Figure 15: Illustration showing quantities appearing in the integral for the case of uniform illumination (Similar to Figure 8).

$A_{\underset{\sim}{r}} \cdot \underset{\sim}{r}_i = 0$ and for a truly distant source:

$A_{\underset{\sim}{r}} = 0$

Replace $\underset{\sim}{r}_i$ by $k\underset{\sim}{r}_i$ and let $k \to \infty$ then:

$$I_{\underset{\sim}{r}} = 0 \ , \ I_{\underset{\sim}{n}} \text{ unchanged}$$
$$E_{\underset{\sim}{r}} = 0 \ , \ E_{\underset{\sim}{n}} \text{ unchanged}$$
$$G_{\underset{\sim}{r}} = (1/r_e)(\hat{\underset{\sim}{r}}_i - G\,\hat{\underset{\sim}{r}}_e) \ , \ G_{\underset{\sim}{n}} = 0$$

In addition choosing the z-axis along $\underset{\sim}{r}_i$ removes further terms.

2. DISTANT CAMERA: (Telephoto lens or the object subtends a small angle at the camera)

Replace $\underset{\sim}{r}_e$ by $k\underset{\sim}{r}_e$ and let $k \to \infty$ then:

$$I_{\underset{\sim}{r}} \text{ and } I_{\underset{\sim}{n}} \text{ unchanged}$$
$$E_{\underset{\sim}{r}} = 0 \ , \ E_{\underset{\sim}{n}} \text{ unchanged}$$
$$G_{\underset{\sim}{r}} = (1/r_i)(\hat{\underset{\sim}{r}}_e - G\,\hat{\underset{\sim}{r}}_i) \ , \ G_{\underset{\sim}{n}} = 0$$

In addition choosing the z-axis along $\underset{\sim}{r}_e$ removes further terms.

3. DISTANT SOURCE AND DISTANT CAMERA:

$$I_{\underset{\sim}{r}} = 0 \ , \ I_{\underset{\sim}{n}} \text{ unchanged}$$
$$E_{\underset{\sim}{r}} = 0 \ , \ E_{\underset{\sim}{n}} \text{ unchanged}$$
$$G_{\underset{\sim}{r}} = 0 \ , \ G_{\underset{\sim}{n}} = 0$$

Most practical situations are an approximation of this case.

4. SOURCE AT THE CAMERA:

$$\underset{\sim}{r}_i = \underset{\sim}{r}_e \qquad I = E \text{ and } G = 1$$
$$I_{\underset{\sim}{r}} = E_{\underset{\sim}{r}} \text{ unchanged}$$
$$I_{\underset{\sim}{n}} = E_{\underset{\sim}{n}} \text{ unchanged}$$
$$G_{\underset{\sim}{r}} = 0 \text{ and } G_{\underset{\sim}{n}} = 0$$

5. DISTANT SOURCE AT DISTANT CAMERA:

$$I_{\underset{\sim}{r}} = E_{\underset{\sim}{r}} = G_{\underset{\sim}{r}} = 0$$
$$I_{\underset{\sim}{n}} = E_{\underset{\sim}{n}} \text{ unchanged}, \quad G_{\underset{\sim}{n}} = 0$$

Choosing the object to be on the z-axis removes further terms. This is the simplest possible case.

6. UNIFORM ILLUMINATION:

Uniform illumination (or an approximation thereof) is fairly common and might at first sight appear not to fit into our framework. This subsection shows the equivalence of uniform illumination to one where a point-source is at the camera and a different reflectivity function obtains.

The integrals here are analogous to the ones obtained for the integrating photometer except that we have constant emittance angle rather than constant incident angle. If the incident light intensity is I_1 per unit area oriented in any direction, then it is easy to show that I_1/π falls per unit solid angle per unit area perpendicular to it. The emitted light per unit solid angle per unit area perpendicular to the emitted ray is thus:

$$I_1 \Psi(e) = I_1 (1/\pi) [\int_0^{2\pi} \int_0^{\pi/2} \phi(i,e,g) * (1/2) * \sin(2i) \, di \, dA \,] / \cos(e)$$

This is the same situation as if we had a source at the camera and a reflectivity function such that:

$$\phi(E,E,1) = \Psi(E)$$

(Except that for uniform illumination a certain amount of self-shadowing can occur for non-convex objects)

2.5 THE FIVE O.D.E.'S FOR THE IMAGE ILLUMINATION EQUATION:

$$F(x, y, z, p, q) = A(\underset{\sim}{r})\ \phi(I,E,G) - b(\underset{\sim}{r}') = 0$$

We know $A(\underset{\sim}{r})$ and $\phi(I,E,G)$, and obtain $b(\underset{\sim}{r}')$ from the image. We need F_x, F_y, F_z, F_p and F_q. Since $\underset{\sim}{r} = (x,y,z)$ and $\underset{\sim}{n} = (-p,-q, 1)$ we can get all of these derivatives from $F_{\underset{\sim}{r}}$ and $F_{\underset{\sim}{n}}$.

$$F_{\underset{\sim}{r}} = A(\underset{\sim}{r})\ \phi_{\underset{\sim}{r}}(I,E,G) + A_{\underset{\sim}{r}}(\underset{\sim}{r})\ \phi(I,E,G) - b_{\underset{\sim}{r}}(\underset{\sim}{r}')$$
$$F_{\underset{\sim}{n}} = A(\underset{\sim}{r})\ \phi_{\underset{\sim}{n}}(I,E,G)$$

Let $\underset{\sim}{a} = (I,E,G)$ then:

$$\phi_{\underset{\sim}{r}} = \phi_{\underset{\sim}{a}}\, a_{\underset{\sim}{r}} \text{ and } \phi_{\underset{\sim}{n}} = \phi_{\underset{\sim}{a}}\, a_{\underset{\sim}{n}}$$

Note that $a_{\underset{\sim}{r}}$ and $a_{\underset{\sim}{n}}$ are 3 by 3 matrices, the rows of which we computed in a previous subsection (2.3.4).

$$\underset{\sim}{\underset{\sim}{a}}_{\underset{\sim}{r}} = \begin{pmatrix} (1/r_i)(\hat{\underset{\sim}{n}} - I\ \hat{\underset{\sim}{r}}_i) \\ (1/r_e)(\hat{\underset{\sim}{n}} - E\ \hat{\underset{\sim}{r}}_e) \\ (1/r_e)(\hat{\underset{\sim}{r}}_i - G\ \hat{\underset{\sim}{r}}_e) + (1/r_i)(\hat{\underset{\sim}{r}}_e - G\ \hat{\underset{\sim}{r}}_i) \end{pmatrix}$$

$$= \begin{pmatrix} 1/r_i & -I/r_i & 0 \\ 1/r_e & 0 & -E/r_e \\ 0 & (1/r_e - G/r_i) & (1/r_i - G/r_e) \end{pmatrix} \begin{pmatrix} \hat{\underset{\sim}{n}} \\ \hat{\underset{\sim}{r}}_i \\ \hat{\underset{\sim}{r}}_e \end{pmatrix}$$

Note that this is the product of two 3 by 3 matrices. Similarly:

$$\underset{\sim}{\underset{\sim}{a}}_{\underset{\sim}{n}} = \begin{pmatrix} (1/n)(\hat{\underset{\sim}{r}}_i - I\ \hat{\underset{\sim}{n}}) \\ (1/n)(\hat{\underset{\sim}{r}}_e - E\ \hat{\underset{\sim}{n}}) \\ \underset{\sim}{0} \end{pmatrix}$$

$$= \begin{pmatrix} -I/n & 1/n & 0 \\ -E/n & 0 & 1/n \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{\underset{\sim}{n}} \\ \hat{\underset{\sim}{r}}_i \\ \hat{\underset{\sim}{r}}_e \end{pmatrix}$$

The calculation of $b_{\underset{\sim}{r}}(\underset{\sim}{r}')$ from $b_{\underset{\sim}{r}'}(\underset{\sim}{r}')$ will be described in section 2.7 .

2.6 CAMERA PROJECTION EQUATIONS:

The projection equations derived here are used in section 2.7 .

So far we have assumed the camera to be at the origin oriented with its optical axis directed along the z-axis and the image-plane x′ and y′ axes parallel to the x and y axes. Moving the camera from the origin introduces only a minor change in the equations. If however the camera is oriented in a different way, some of the equations become more complicated.

Let R be the orthonormal 3 by 3 matrix which takes the z-axis into the optical axis and the x and y axes into the x′ and y′ axes. Then:

$$\underset{\sim}{r}' = (R\ \underset{\sim}{r})\ \frac{f}{(R\ \underset{\sim}{r}).\hat{\underset{\sim}{z}}}$$

where $\hat{\underset{\sim}{z}} = (0,0,1)$ is the unit vector along the z-axis in the image coordinate system. $\underset{\sim}{r}'$ is the vector from the exit-pupil to the image point in the same coordinate system.

If two images are taken with the camera oriented differently, the area recorded in both images will be spatially distorted

only. That is, a simple transformation will take the one image into the other.

$$\underset{\sim}{r}_2' = \frac{(R_2 R_1^{-1}\ \underset{\sim}{r}_1')}{(R_2 R_1^{-1}\ \underset{\sim}{r}_1').\hat{\underset{\sim}{z}}} * \frac{f_2}{f_1}$$

Where R_1 and R_2 are the two rotation matrices and f_1 and f_2 the corresponding exit pupil to image plane distances. This transformation is useful if one wishes to orient the optical axis along $\underset{\sim}{r}_i$ or $\underset{\sim}{r}_e$ (to simplify the equations for the derivatives).

2.7 OBTAINING INTENSITY GRADIENTS:

To evaluate the derivative $F_{\underset{\sim}{r}}$ (section 2.5) we need $b_{\underset{\sim}{r}}(\underset{\sim}{r}')$.

$$b_{\underset{\sim}{r}}(\underset{\sim}{r}') = b_{\underset{\sim}{r}'} \cdot \underset{\sim}{r}'_{\underset{\sim}{r}}$$

$$\underset{\sim}{r}'_{\underset{\sim}{r}} = R\underset{\sim}{r}_{\underset{\sim}{r}} \frac{f}{(R\underset{\sim}{r}).\hat{\underset{\sim}{z}}} - (R\underset{\sim}{r}) \frac{f\ R\underset{\sim}{r}_{\underset{\sim}{r}}.\hat{\underset{\sim}{z}}}{(R\underset{\sim}{r}.\hat{\underset{\sim}{z}})^2}$$

$$= \frac{f}{(R\underset{\sim}{r}).\hat{\underset{\sim}{z}}} \left(R - \frac{(R\underset{\sim}{r})(R\hat{\underset{\sim}{z}})}{(R\underset{\sim}{r}).\hat{\underset{\sim}{z}}}\right)$$

In the simple case that the camera is oriented properly:

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\underset{\sim}{r}' = \frac{f}{\underset{\sim}{r}\cdot\hat{\underset{\sim}{k}}}\left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{\hat{\underset{\sim}{r}}\,\hat{\underset{\sim}{k}}}{\underset{\sim}{r}\cdot\hat{\underset{\sim}{k}}}\right]$$

$$= \frac{f}{z}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{f}{z^2}\begin{pmatrix} 0 & 0 & x \\ 0 & 0 & y \\ 0 & 0 & z \end{pmatrix}$$

$$= \frac{f}{z}\begin{pmatrix} 1 & 0 & -x/z \\ 0 & 1 & -y/z \\ 0 & 0 & 0 \end{pmatrix}$$

Written out in full we have:

$$(b_x, b_y, b_z) = (f/z)[b_{x'}, b_{y'}, -((x/z)b_{x'} + (y/z)b_{y'})]$$

$b_{x'}$ and $b_{y'}$ are measured directly from the image.

Since the intensities measured from the image do not locally determine the normal, one might well ask what, roughly, such measurements do determine. The components of the gradient of the intensity are related to the second derivatives of the distance to the surface, while the intensity itself is related to the magnitude of the first derivatives. This relationship becomes exact for the case of a distant source at a distant camera (section 2.5, case 5; see also 3.1.2).

It should be noted that the equation for $F_{\underset{\sim}{r}}$ (section 2.5)

also involves $A_{\underset{\sim}{r}}$. Usually A is fairly constant over the area of the object recorded in the image, or at least satisfies a simple inverse-square equation.

If $A = (r_\ell / r_i)^2$, then $A_{\underset{\sim}{r}} = -2(r_\ell^2 / r_i^4)\underset{\sim}{r}_i$.

Where $\underset{\sim}{r}_i$ is the incident vector, and r_ℓ is the length of the incident vector to the singular point.

2.8 OBTAINING INITIAL CONDITIONS:

It would be a great disadvantage if one always required an initial curve to start the solution from. Fortunately it is usually possible to calculate some initial curve if one makes some assumptions about the surface and uses the special points where the reflectivity uniquely determines the local normal - these points will be called singular points.

2.8.1 USE OF THE SINGULAR POINTS:

The singular points are the brightest or the darkest points (depending on the reflectivity function). At all other points the normal cannot be locally determined. The singular

points are points corresponding to values of i and e for which the reflectivity is a unique global maximum or minimum. These may be either extrema in the calculus sense or at the limiting values of the angles.

This method cannot be used if the surface does not contain a surface element oriented in this special direction. The points are found by looking for the brightest (or darkest) points in the image.

All we still need to know then is the distance of this point from the camera, but since one is usually only interested in relative distances this is not a serious restriction. Unfortunately it will be found that the solution will not move from these singular points, i.e. $\dot{x}' = \dot{y}' = 0$. This is an indication that the algorithm needs to be informed about which way the surface is curved (convex or concave). To make this more concrete assume we have a distant source and can thus calculate G at each image point.

2.8.2 THE SOLUTION WILL NOT MOVE FROM A SINGULAR POINT:

Consider the variation of ϕ with E first:

1. If the extremum occurs for $0<E<1$ then $\phi_E = 0$.

2. If the extremum occurs for $E = 1$ then $\hat{\underset{\sim}{n}} = \hat{\underset{\sim}{r}}_e$ and hence $E_{\underset{\sim}{n}} = (1/n)(\hat{\underset{\sim}{r}}_e - E\,\hat{\underset{\sim}{n}}) = 0$.

3. If the extremum occurs for $E = 0$ then $\hat{\underset{\sim}{n}}.\hat{\underset{\sim}{r}}_e = 0$ and $E_{\underset{\sim}{n}} = (1/n)\hat{\underset{\sim}{r}}_e$. That is, $xp+yp-z=0$ and $E_p = (1/nr)x$ and $E_q = (1/nr)y$.

$$\dot{x} = \phi_E(1/nr)x \quad \text{and} \quad \dot{y} = \phi_E(1/nr)y$$
$$\dot{z} = \phi_E(1/nr)(px+qy) = \phi_E(1/nr)z$$

In case 1 and 2 we have $\phi_E E_p$ and $\phi_E E_q = 0$.

Now consider the variation of ϕ with I:

1. If the extremum occurs for $0<I<1$ then $\phi_I = 0$.

2. If the extremum occurs for $I = 1$ then $\hat{\underset{\sim}{n}} = \hat{\underset{\sim}{r}}_i$ and hence $I_{\underset{\sim}{n}} = (1/n)(\hat{\underset{\sim}{r}}_i - I\,\hat{\underset{\sim}{n}}) = 0$.

3. If the extremum occurs for $I = 0$ then $\hat{\underset{\sim}{n}} \cdot \hat{\underset{\sim}{r}}_i = 0$ and $I_{\underset{\sim}{n}} = (1/n)\hat{\underset{\sim}{r}}_i$. That is $(x-x_o)p + (y-y_o)q - (z-z_o) = 0$ and $I_p = (1/nr_i)(x-x_o)$ and $I_q = (1/nr_i)(y-y_o)$.

$$\dot{x} = \oint (1/nr_i)(x-x_o), \qquad \dot{y} = \oint (1/nr_i)(y-y_o)$$
$$\dot{z} = \oint (1/nr_i)((x-x_o)p + (y-y_o)q)$$
$$= \oint (1/nr_i)(z-z_o)$$

In case 1 and 2 we have $\oint_I I_p$ and $\oint_I I_q = 0$.

Now $\dot{x} = F_p$ and $\dot{y} = F_q$ and $\dot{z} = pF_p + qF_q$.

$$F_{\underset{\sim}{n}} = \Lambda(\underset{\sim}{r})\ \oint(I,E,G)$$
$$\oint_p = \oint_I I_p + \oint_E E_p$$
$$\oint_q = \oint_I I_q + \oint_E E_q$$

So in all combinations of cases 1 and 2 for E and cases 1 and 2 for I we find $\dot{x} = \dot{y} = 0$ and hence also $\dot{z} = 0$, therefore:

$$\dot{x}' = \dot{y}' = 0$$

That is, the projection of the solution point into the image is not moving as the parameter s is changed.

In the case $E = 0$ we find that even though $\dot{x}$ and $\dot{y}$ may be non-zero, $\dot{x}'$ and $\dot{y}' = 0$.

All that remains is the case $I = 0$. Here too $\dot{x}' = \dot{y}' = 0$, when the source is at the camera or if E is undetermined. We have found no paints with an extremum for $I = 0$ where E was determined (i.e. the global extremum was not unique).

2.8.3 GETTING THE INITIAL CURVE FROM A SINGULAR POINT:

If the surface is convex (or concave) at the singular point and we have a guess at the radius of curvature (from the overall size of the object for example) we can get around the problem of singular points by constructing small spherical caps on them. Difficulties will be encountered if this point happens to be a saddle point (The presence of a saddle point however usually indicates that other singular points exist where the surface is either convex or concave).

Let $\underset{\sim}{S}$ be the vector from the camera to the singular point (found from its known image coordinates and its distance from the camera). R is the estimated radius of curvature and ρ the distance we decide to step away from the singular point (determined in practice by considerations of uncertainty in

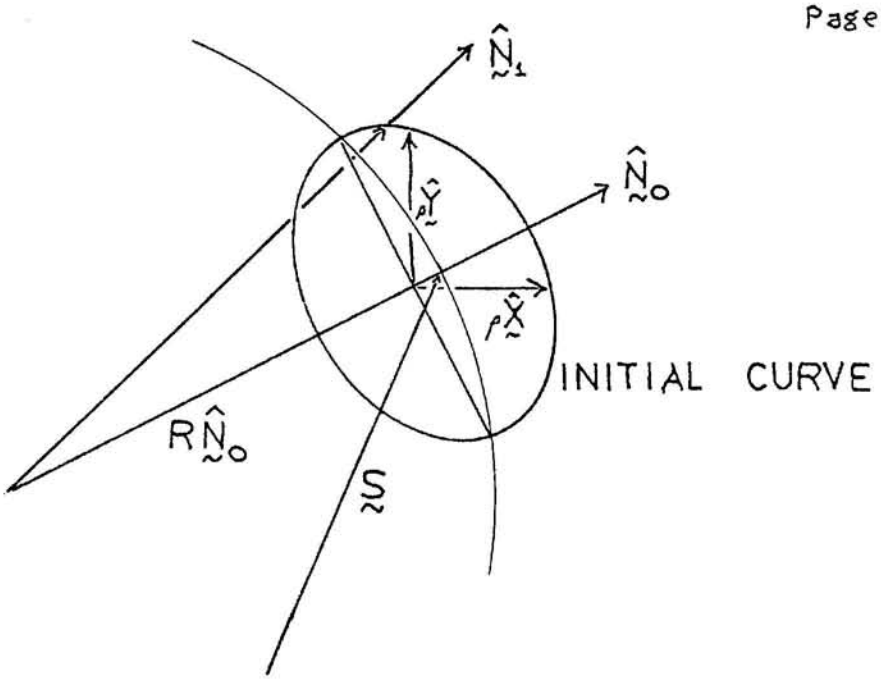


Figure 16: Construction of the initial curve near a singular point.

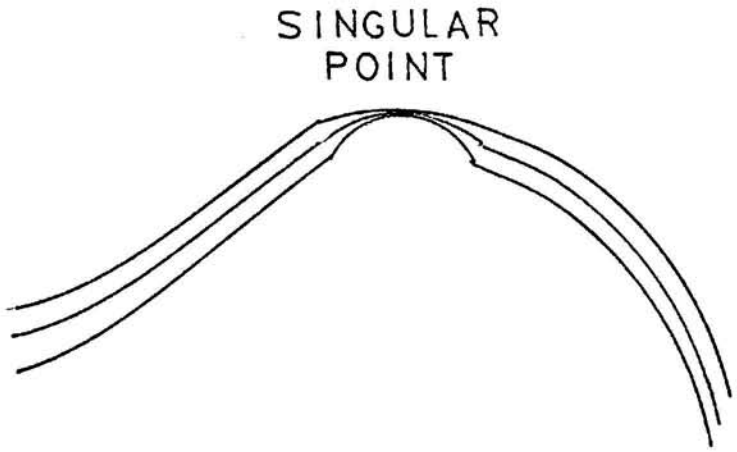


Figure 17: Illustration portraying three solutions obtained for varying initial radius of curvature - showing the small effect which errors in the initial curve have on the solution.

the position of the singular point and the desired detail in the solution). The known normal at the singular point is $\hat{\underset{\sim}{N}}_o$. We construct a spherical cap with center $\underset{\sim}{S} - R\hat{\underset{\sim}{N}}_o$.

Let $$R_1^2 = R^2 - r^2$$

$$\underset{\sim}{S}_1 = \underset{\sim}{S} + (R_1 - R)\hat{\underset{\sim}{N}}_o$$

Let $$\underset{\sim}{X} = \hat{\underset{\sim}{y}} \times \hat{\underset{\sim}{N}}_o \quad \text{where} \quad \hat{\underset{\sim}{y}} = (0,1,0)$$

$$\underset{\sim}{Y} = \hat{\underset{\sim}{N}}_o \times \hat{\underset{\sim}{X}}$$

$$\underset{\sim}{T}(t) = \rho[\hat{\underset{\sim}{X}} \cos(2\pi t) + \hat{\underset{\sim}{Y}} \sin(2\pi t)] \qquad 0 \leq t < 1$$

Points on the initial circle are then given by

$$\underset{\sim}{S}_1 + \underset{\sim}{T}(t)$$

We also need an initial guess at p and q, so we construct $\underset{\sim}{N}_1$, (an outward normal):

$$\underset{\sim}{N}_1(t) = R_1 \hat{\underset{\sim}{N}}_o + \underset{\sim}{T}(t)$$

The requirement for an initial guess at the radius of curvature is not as restrictive as it might seem, since the required accuracy is extremely low. This is because is usually very much smaller than R, and hence a change in R affects the position of the initial curve very little. Even more importantly the values derived for p and q need not be

accurate since they are only used as a first guess in an iterative method of finding p and q on the initial curve before starting the solution.

2.9 NON-POINT SOURCES:

Uniform sources have already been dealt with. Perhaps the easiest other case is a circularly symmetric source at a distance large compared to the dimensions of the object.

2.9.1 CIRCULARLY SYMMETRIC SOURCES:

Distant circularly symmetric sources can be replaced by a point source after modifying the reflectivity function. One merely convolves the reflectivity function with the spread function of the source (a bit of spherical trigonometry is involved here). Strictly speaking one should perform the same operation with the entrance pupil of the camera since it too subtends a finite angle at the object and accepts a bundle of light-rays. Since ϕ is smooth (except at $I = 0$ and $I = 1$) it will be changed very little except at these points. The main change will be that ϕ does not tend to 0 as I tends to 0, but rather for some negative value of I. Also the

specular component will be more smeared out.

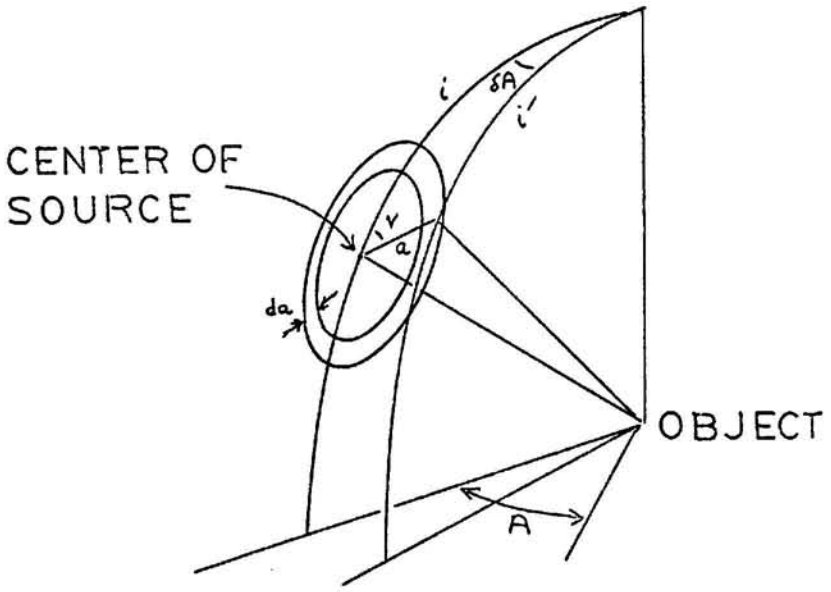


Figure 18: Illustration of circularly symmetric source and quantities used in the convolution.

Let the source intensity be I(a) per unit solid angle at the angle a from its center when viewed from the object. Then the new reflectivity function $\phi'(I,E,G)$ is:

$$\phi'(I,E,G)=\int_0^{2\pi}\int_0^{a_0} I(a)\ \phi(I',E,G')\ a\ da\ dv \Big/ \int_0^{2\pi}\int_0^{a_0} I(a)\ a\ da\ dv$$

Where a_0 is the total angular diameter of the source.

And $\cos(A) = (\cos(g)-\cos(i)\cos(e))/(\sin(i)\sin(e))$

$\cos(i') = \cos(i)\cos(a) + \sin(i)\sin(a)\cos(v)$

$\sin(\delta A) = \sin(i')\sin(a)/\sin(v)$

$\cos(g') = \cos(A+\delta A)\sin(i')\sin(e) + \cos(i')\cos(e)$

2.9.2 MULTIPLE SOURCES:

When the source distribution is not easily treated as above one can introduce a different A_k for each source and replace the main equation by:

$$\sum_k A_k(\underset{\sim}{r})\ \phi(I_k, E, G_k) = b(\underset{\sim}{r}')$$

Difficulties in finding initial conditions will be encountered with multiple sources unless they are of special kinds (e.g. a point source and a uniform source).

2.10 TYPES OF EDGES:

Several kinds of edges appear in an image - each with its own properties and problems for our algorithm:

1. Overlap - (occlusion of one object by another) discontinuity in z. The program must detect this or it will erroneously continue a solution across such an edge.

2. Joints - (angular edges on an object) discontinuities in the derivatives of z. One cannot continue p and q across such an edge. It is possible however to use the position of the edge as a new initial curve. This and the previous condition can be detected as a step in the intensity distribution or from a highlight on the edge.

3. View edges - special case of *1.* , where no joint appears, i.e. the surface is smooth and E tends to *0* as we approach it. This is easily detected by the program during the calculation of the solution.

4. Shadow edges - here I tends to *0* as we approach the edge and again the program can easily detect this.

5. Other edge of shadow - if the shadow was bridged this edge may serve as a new initial curve.

6. Ambiguity edges - some are lines of aggregation of singular points (on which $\lambda \rightarrow \infty$). The characteristics will not cross an ambiguity edge (see section *3.1.3*).

2.11 SHADOWS AND SELF-ILLUMINATION:

If the single source is not at the camera, shadows will appear. Solutions can be carried across shadows since the position of the source is known and one can construct a ray through the last illuminated point and trace it until it meets another illuminated region. Only the coordinates and not the local gradient of this new point will be known. It is necessary to carry this operation out for all characteristics entering the shadow, producing a new initial curve at the other edge of the shadow where we can restart the solution. In practice care has to be taken because of noisyness of the solution.

Self-illumination is a difficult problem to deal with unless the object is convex or its albedo is low (less than 0.4). An estimate of the effect of self-illumination can be obtained from a consideration of two semi-infinite matt planes joined at right angles. These are illuminated from a very great distance and such that the incident rays make an angle a with one of the planes. Let the reflectivity of the surface obey lamberts law and the fraction of the incident light reflected be k. Contrast between two intensities I_1 and I_2 is usually defined to be:

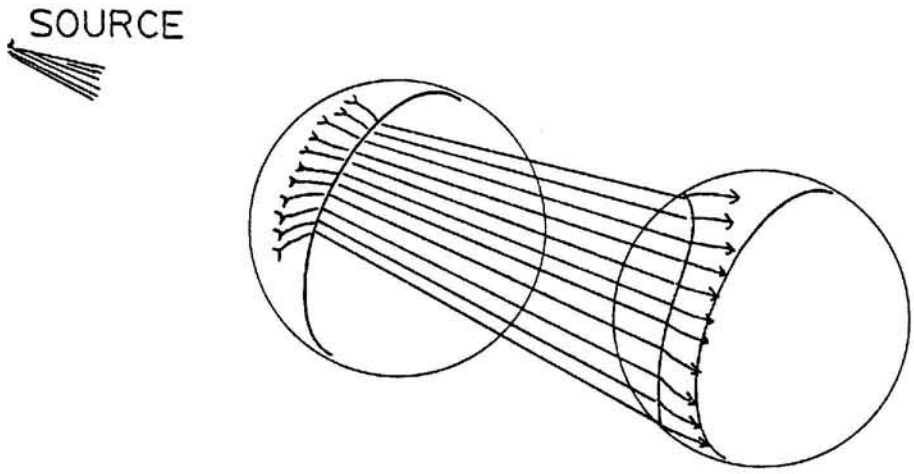


Figure 19: Bridging a shadow.

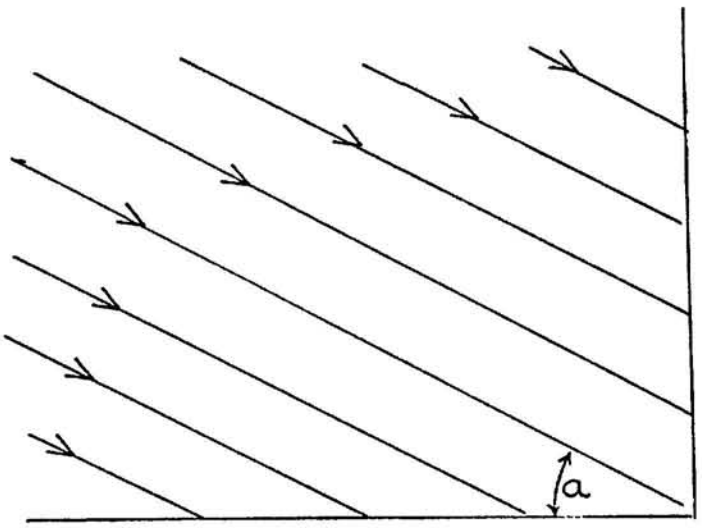


Figure 20: Two semi-infinite planes joined at right angles. Used in the study of self-illumination.

$$C = \left| \frac{I_1 - I_2}{I_1 + I_2} \right|$$

If we ignore light reflected more than once, we find the contrast between the two planes to be:

$$C_1 = \tan(a - \pi/4)$$

While if the self-illumination is taken into account we get:

$$C_2 = \frac{2 - k}{2 + k} \tan(a - \pi/4)$$

Contrast is thus reduced by a factor of $(2 - k)/(2 + k)$. This factor varies from *1/3* to *1* as k varies from *1* to *0*.

Note: the rest of chapter *2* contains some miscellaneous items that did not fit in elsewhere.

2.12 THE INVERSE PROBLEM - GENERATING HALF-TONE IMAGES:

The inverse problem of producing images of a specified scene with shading and shadows is vastly different from the method of shape-from-shading. Most programs written for this purpose can be used for objects bounded by planes only. The main issues of optimization of the calculation of which

surfaces are visible to the source and camera respectively have been dealt with in some detail in recent work [8]. Although the two problems are inverses of one another, the methods used are quite different.

An interesting problem of a mathematical nature (and incidentally with application to cutting wood-cuts) is that of producing curved lines in a plane such that the density of lines is proportional to the shading in the image of some real or imagined object. Preferrably one would like as small a number of 'unneccesary' breaks in the lines as possible, i.e. the lines should either close on themselves or leave the image. Another restriction one might apply is that the lines should not cross (When producing wood-cuts one would most likely also reflect some of the surface texture in the choice of lines).

For a special case, a solution is immediately at hand. This is the case where we have a distant camera at a distant source (section 2.4, case 5; see also 3.1.2) and a reflectivity function ϕ such that:

$$\phi(I,I,1) = I = 1/\sqrt{1+p^2+q^2}$$

Here the contour lines give a solution, with no crossing lines and no 'unneccesary' breaks. One of the most attractive feature of contour maps is perhaps just this fact that they provide some shading information.

2.13 HUMAN PERFORMANCE WITH MONOCULAR PICTURES:

Jugding by the popularity of monocular pictures of people and other smooth objects, humans are good at interpreting shading information. Since they use the same basic information as our shape-from-shading algorithm we expect to find similar short-comings (see section on facial make-up for example). Supposing the human visual system does not use the shading information in simple heuristic ways only, one might expect that the perception system 'solves' the equations or a much simplified form of them. Since this cannot be done locally (the way some portions of an edge-finding process might work) it is difficult to suggest an elegant and simple mechanism and a place to look for it. Presumably it would have to involve computational waves travelling outward from the singular points.

2.14 ERRORS AND INCONSISTENCIES:

It is difficult to estimate analytically the error in the solution because the equations are so non-linear. ϕ, b, and A cannot be measured to better than 5 or 10% accuracy and numerous practical problems such as non-uniform sensitivity of the sensor have to be taken care of.

Only a simple error analysis can be presented here. Suppose we wish to determine the effect of varying inclinations on how a given error in the input data (intensity in the image) relates to errors in the coordinates determined on the characteristics. We need to determine the rate of change of p w.r.t. b. Consider a particularly simple case, that of a distant source at a distant camera (As has been mentioned previously and will be demonstrated in section 3.1, the equations for this case are particularly simple). Next assume that one of the gradient components, q say, is 0.

We have $b/A = \phi(I) = \phi(1/\sqrt{1+p^2+q^2}) = \phi(1/\sqrt{1+p^2})$

Then $p = \sqrt{1/(\phi^{-1}(b/A))^2 - 1}$

We need to differentiate p w.r.t. the ratio $l = b/A$.

$$p_l = \frac{-\phi_l^{-1}(l)}{(\phi^{-1}(l))^2\sqrt{1-(\phi^{-1}(l))^2}}$$

For both $I \rightarrow 0$ and $I \longrightarrow 1$, the error in p becomes very large for a given error in l (since in the first case $\phi^{-1}(l) \longrightarrow 0$ and in the second case $\phi^{-1}(l) \longrightarrow 1$). This is not very surprising since in the first case we are looking perpendicularly down on the surface and I will vary very slowly with p, while in the second case we have near tangential incidence and small changes in the angle of incidence (and hence also I) will correspond to large changes in p.

We note that in this rather special case, the error contribution to the solution is large in some areas, while being small in others where the incident angle is not to close to 0 or $\pi/2$. The actual error will also depend on ϕ^{-1} and the error in measuring b/A. In a case with less restricted lighting conditions the relationship between the inclination of the surface and the error-rate will be more complex.

We considered the derivative of p w.r.t. l, since it is the integral of the error in p which constitutes the error in z for any one characteristic.

$$e(s) = \int_0^s \delta p(s)\, ds = \int_0^s p\ \delta l(s)\, ds$$

Where $e(s)$ is the error in z for a given characteristic as a function of arc-distance from the singular point, $\delta p(s)$ is the error in p and $\delta l(s)$ is the error in l.

In this context one may also want to discuss inconsistencies in the solution. If either the lighting conditions or the reflectivity function are incorrectly specified, an incorrect shape will be calculated. The shape determined may or may not violate the requirement of smoothness. If the calculated shape is not smooth it can be concluded that the solution (at least in some region) is incorrect, and that the given source position or the given reflectivity function are incorrect. It is easy to give examples of the case where false assumption will lead to a smooth solution, as well as those where we obtain solutions with discontinuities.

For simplicity consider a flat, inclined surface ($z = x$). The characteristics will be straight lines in this plane, parallel to the x-z plane. Modifying the reflectivity of the surface to be increasingly darker with increasing x, we obtain a new solution which contains characteristics, again parallel to the x-z plane, but curving toward large z for

large x. This solution is smooth and contains no indication of an error.

If now we apply instead a surface coating which is normal for positive y and darker for negative y, we obtain a solution in which the inconsistency is apparent. The characteristics in the solution for negative y are more inclined than those for positive y, and a discontinuity exists at y = 0.

Using this kind of approach one could determine which kind of surface markings are noticeable by an observer (i.e. lead to inconsistencies in the solution) and those which merely alter the apparent shape.

2.15 WHAT ARE LIKELY SOURCE DISTRIBUTIONS?

Since the complexity of the algorithm presented here increases with the complexity of the light-source distribution and since we only know how to bridge shadows cast by one source, it is important to know which light-source distributions occur in practice. First one notes that the situations found difficult by humans are almost certainly going to give difficulties to our algorithm. For example, when two sources cast shadows (such as on a road lighted by

widely spaced street-lamps) the shape of unfamiliar objects becomes difficult to ascertain because of the crossed shadows. If the incident intensity varies greatly from one image area to another (such as in a lightly wooded forest) the tangle of lighted and dark areas makes perception more difficult. On the other hand one would expect 'natural' conditions to be particularly easy. That is, one point source somewhat above the observer (the sun) combined with a very diffuse (almost uniform) source (the sky). The diffuse source will not throw sharp shadows of its own. The absence of either of the two sources makes vision only slightly more difficult.

2.15.1 RELEVANCE TO PHOTOGRAPHY AND GRAPHICS:

One would expect photographers to have something to contribute to this subject and introductory booklets on artificial light photography confirm the above conclusions. The beginner is advised to use a number of lights with different characteristics as follows (Phrases of inexact meaning will be placed in quotes):

1. The main light - The ideal main light is a large spot light approximating the effect of the sun. It is usually

placed 45 degrees above and 45 degrees to the side of the subject. Its purpose is to establish the 'form of the subject' and fix the ratio of lighted to dark areas. The exact ratio is not important but the position of the source should result in good shading (which increases as the source is moved further from the camera) without too much shadow area (in which detail is more difficult to perceive).

2. The fill-in light (or axial light) - Its purpose is to lighten slightly the shadows cast by the main light and approximates the effect of the sky. It is placed near the camera to prevent it from casting its own shadows and to simulate the effect of uniform lighting (see an earlier discussion of uniform illumination, section 2.4.6). The appearance of shadows within shadows is considered extremely 'ugly' and should be avoided since it makes the picture more difficult to interpret. The ratio of fill-in light intensity to main light intensity is usually chosen to be about 1 to 3.

In addition a number of small sources may be used for extra effects:

3. The accent light - Its purpose is to enliven the rendering by adding highlights and 'sparkle'. It should be a small collimated source which can be directed to illuminate small sections of the subject. It is placed behind and to the side of the subject so that it cannot cast shadows of its own. This light can add catchlights (specular reflections such as on eyes or metal objects) and bright outlines (particularly on hair).

4. The background light - Its purpose is to 'separate' the subject from the background. It illuminates the background only, such that the intensity reflected by the subject will nowhere match that of the background. This ensures that the two can be easily 'separated' - i.e. the edge between them will be visible.

Other hints are that too many lights spoil the effect, having the main-light at the camera creates a 'flat' image, shadows crossing edges on the subject are to be avoided and that light parts of the image draw the attention of the viewer. It is interesting to note how much of what is vaguely formulated in these introductions to photography can be understood from the point of view of shading.

2.16 DETERMINING SHAPE FROM TEXTURE GRADIENTS:

A problem related to that of determining shape using shading is that of determining shape from the depth-cue of texture gradients. A textured surface will produce an image in which the texture is distorted in a way reflecting both the direction and and the amount of the inclination of the surface. An image of a tilted surface with a random dot-pattern for example will be compressed in one direction (the average distance between dots is decreased) by an amount proportional to the inclination of the surface. Both direction and magnitude of the gradient can thus be determined - except for a two-way ambiguity.

In practice it may not always be easy to determine such texture gradients reliably because of low resolution of the imaging device and scatter, causing a reduction in contrast. Some simple textures may be handled by simple counting or distance measurements as suggested above, while more complicated textures (e.g. a plastered wall) will need more sophisticated techniques, such as two-dimensional correlation (best obtained using the fast-fourier-transform). Some experimentation with this technique showed promise, but did not supply very reliable gradients and the method was slow.

The next problem is how to obtain the shape from the texture gradients. Starting at some point (whose distance from the camera we assume known), we use some external knowledge to resolve the two-way ambiguity. We can now take a small step in any direction and find the gradient at this new point. Continuing in this way we trace out some curve on the surface of the object (somewhat analagous to the characteristics in the shape-from-shading method, except that here the curve is quite arbitrary).

Let s be the arc-distance along the curve, z_o the distance to the initial point, and p and q the components of the gradient, then:

$$z(s) = z_o + \int_o^s (p,q) \cdot \underset{\sim}{ds}$$

If one takes small enough steps, one can continue to resolve the ambiguity at each step by using the assumption of smoothness. This can be done until we meet a point where the gradient is zero. To continue past such a point would require some external knowledge to again resolve the two-way ambiguity. An aggregation of points with zero inclination can form an ambiguity edge which cannot be crossed.

Clearly we can reach a given point through many paths from the initial point. This allows us some error checking, but there certainly are better ways of making use of the excess information. For that is what we have, since we know from the solution to the shape-from-shading that only one value is required at each point for the determination of the shape, while we here have two (the components of the gradient). Most commonly when faced with such an excess of information on can make use of some least-squares technique to improve the accuracy. Perhaps a relaxation method on a grid would be useful (The grid need not be rectangular).

3. PRACTICAL APPLICATION:

3.1 THE SCANNING ELECTRON MICROSCOPE:

This chapter deals with a few practical applications in which the equations simplify considerably.

3.1.1 DESCRIPTION OF THE SCANNING ELECTRON MICROSCOPE:

This device uses an electron beam which is focused and deflected much like the beam of a cathode ray tube and impinges on a specimen in an evacuated chamber [11]. The narrow ray penetrates into the specimen for some distance, creating secondary electrons along its path (a small number of electrons are reflected at the surface). The depth of penetration, the spread and the number of secondary electrons are all functions of the material of that portion of the specimen. The number of secondary electrons which reach the vacuum through the surface will depend strongly on the inclination of the surface w.r.t. the beam, being least when it is perpendicular.

These relatively slow secondary electrons are then attracted by a positively charged grid and impinge on a phospor-coated

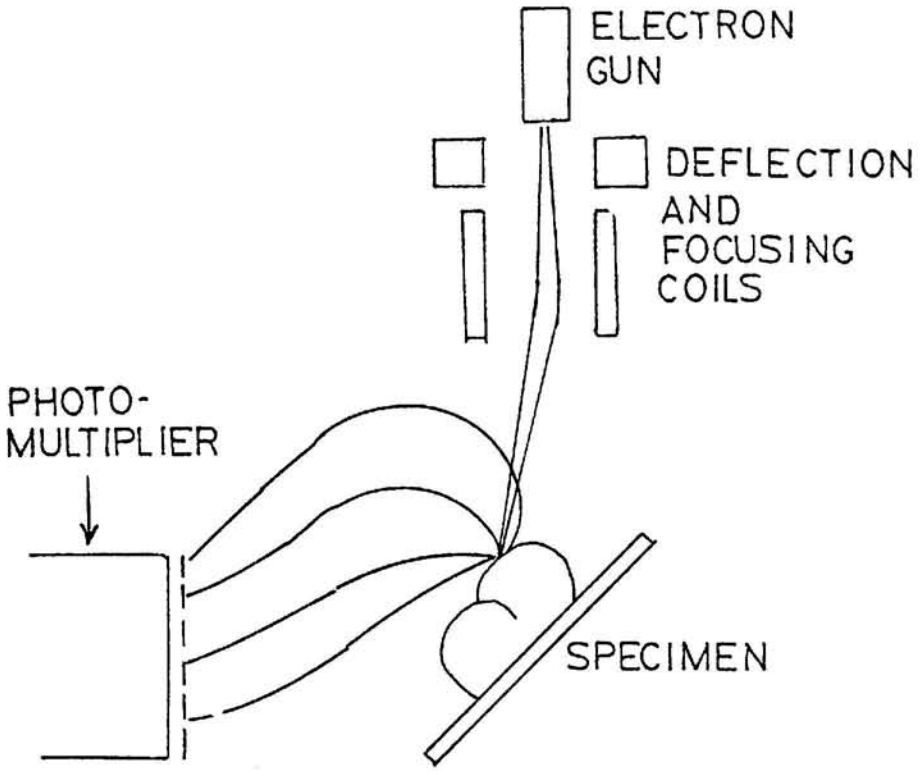


Figure 21: Sketch of a scanning electron microscope.

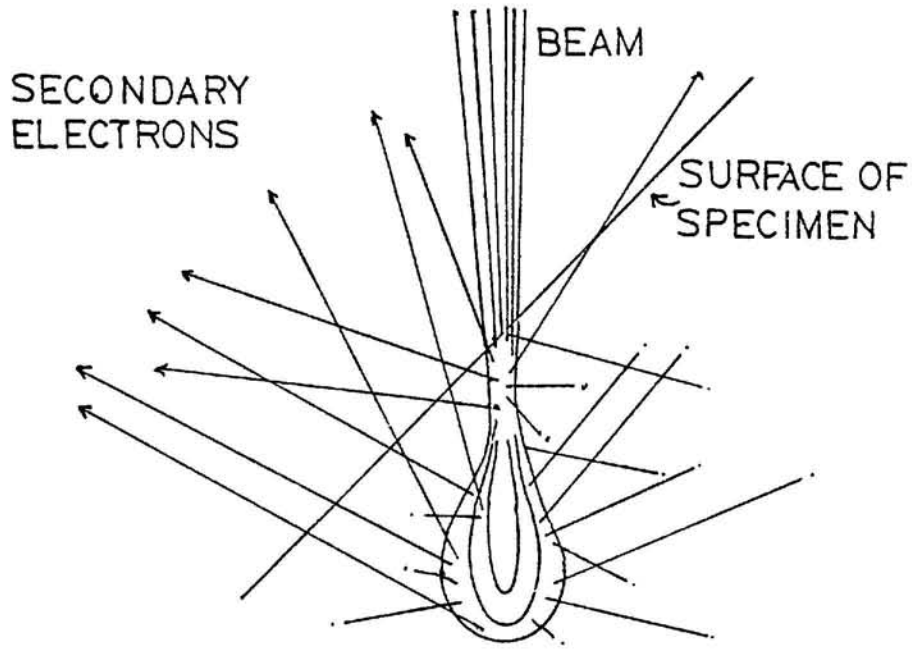


Figure 22: Detail of electron beam impinging on specimen.

photomultiplier. In this way a current is generated proportional to the number of secondary electrons escaping the specimen. There are other modes of operation which do not however interest us here. The output is used to modulate the intensity of the beam in a cathode ray tube while both beams are scanned synchronously in a T.V. like raster. The image created exhibits shading and is remarkably easy to interpret topgraphically. This is quite unlike the normal use of optical or transmission electron microscopes which portray density and thickness.

The magnification is easily increased by decreasing the deflection in the microscope. The resolution is poor compared to the transmission electron microscope because of the spread of the beam as it enters the specimen, but the depth of field is much better than that of an optical microscope because of the very narrow beam (extremely high f-number). The higher field gradient on edges causes these to be outlined more brightly. This artifact, while appealing to people, may be a problem in the implementation of a computer algorithm for finding the shape.

Often the final analysis does not involve exact determination of the shape or two stereo-images can be used, but there propably are also important cases where the shape must be

determined and the stereoscopic method is not applicable. This may be because at the magnification used the specimen appears smooth without significant surface detail or because it is difficult to line up the second image. Since the equations for this case turn out to be so simple it should be rewarding to tie a scanning electron microscope directly into a small computer.

3.1.2 EQUATIONS FOR THE SCANNING ELECTRON MICROSCOPE:

A little thought shows that this is analogous to the case where the source is at the camera (or equivalently we have uniform illumination); for one thing, no shadows appear. Next we note that at all but the lowest magnifications the projection is near-orthogonal. Because of these two effects the five O.D.E.'s simplify considerably:

$$\dot{x} = F_p, \quad \dot{y} = F_q, \quad \dot{z} = pF_p + qF_q$$
$$\dot{p} = -F_x - pF_z \quad \text{and} \quad \dot{q} = -F_y - qF_z$$

Now $\underset{\sim}{F_n} = A\,\phi_I \underset{\sim}{I_n}$ and $\underset{\sim}{F_r} = -\underset{\sim}{b_r}$

$$I = \underset{\sim}{n}\cdot\hat{\underset{\sim}{z}}/n = 1/n \quad (\text{where } \underset{\sim}{n} = (-p,-q,1))$$

$$I_{\underset{\sim}{n}} = (1/n)(\hat{\underset{\sim}{z}} - I\,\hat{\underset{\sim}{n}}) = (\hat{\underset{\sim}{z}}/n) - (1/n^3)\hat{\underset{\sim}{n}}$$

$$I_p = (1/n^3)p \quad \text{and} \quad I_q = (1/n^3)q$$

Hence: $\dot{x} = F_p = (A\,\phi_I/n^3)p, \quad \dot{y} = F_q = (A\,\phi_I/n^3)q$

$$\dot{z} = (A\,\phi_I/n^3)(p^2 + q^2)$$

$$\dot{p} = -b_x \quad \text{and} \quad \dot{q} = -b_y$$

If $\phi_I \neq 0$ everywhere, we can change to a new measure s along the characteristic by multiplying all equations by $\lambda = n^3/(A\,\phi_I)$ and we get:

$$\dot{x} = p, \quad \dot{y} = q, \quad \dot{z} = p^2 + q^2$$

$$\dot{p} = b_x(n^3/(A\,\phi_I)), \quad q = b_y(n^3/(A\,\phi_I))$$

This extremely simple case thus has characteristics which are curves of steepest descent (or ascent). Also note that the equation for z does not couple back into the system of equations (due to the orthogonal projection) thus increasing accuracy. The equations happen to be very similar to the eikonel equations for the paths of light-rays in refractive media. It may be possible to find ready-made solutions to some special cases by using this analogy.

We assumed that $\phi_I \neq 0$; this is equivalent to assuming that an inverse exists which allows us to find I from a measurement of the image intensity:

$$\Psi(\phi(I,I,1)) = I$$

Let $$\xi(x) = (1-\psi^2(x))/(2\psi(x))$$

Then $$\xi(\phi(I,I,1)) = (1/2)*(p^2 + q^2)$$

So we can find at each point the magnitude, but not the direction of the local gradient. This is very different from the method of determining shape from texture gradients (section 2.16), where we can locally determine the gradient except for a two-way ambiguity.

3.1.3 AMBIGUITIES AND AMBIGUITY EDGES:

This is an easy enough example to study ambiguities. Consider the two surfaces:

$$z = z + x^3 \quad , \quad z = z + |x|^3$$

Clearly they cannot be distinguished from monocular views since their gradient magnitudes are identical: i.e. they produce identical intensity distributions in the image. This

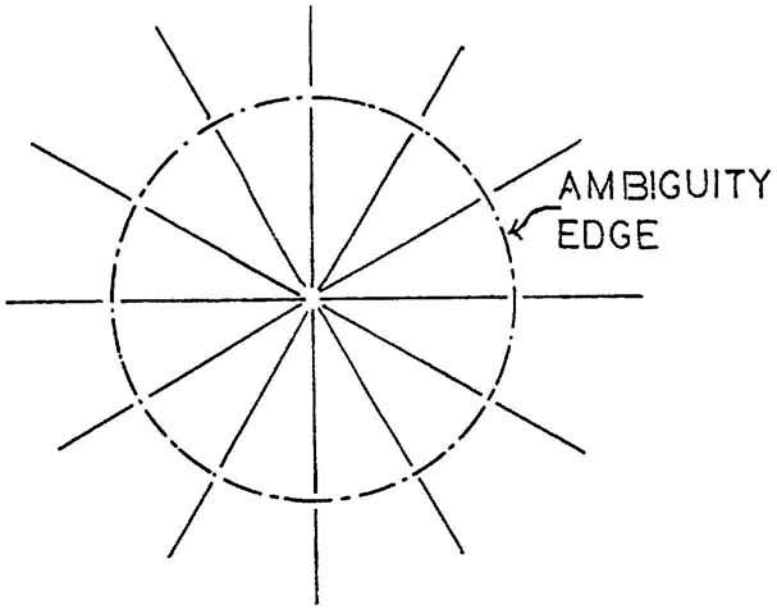


Figure 23: A locally determined ambiguity edge.
$f = 1/\ (x^2 + y^2 - 1)^2$

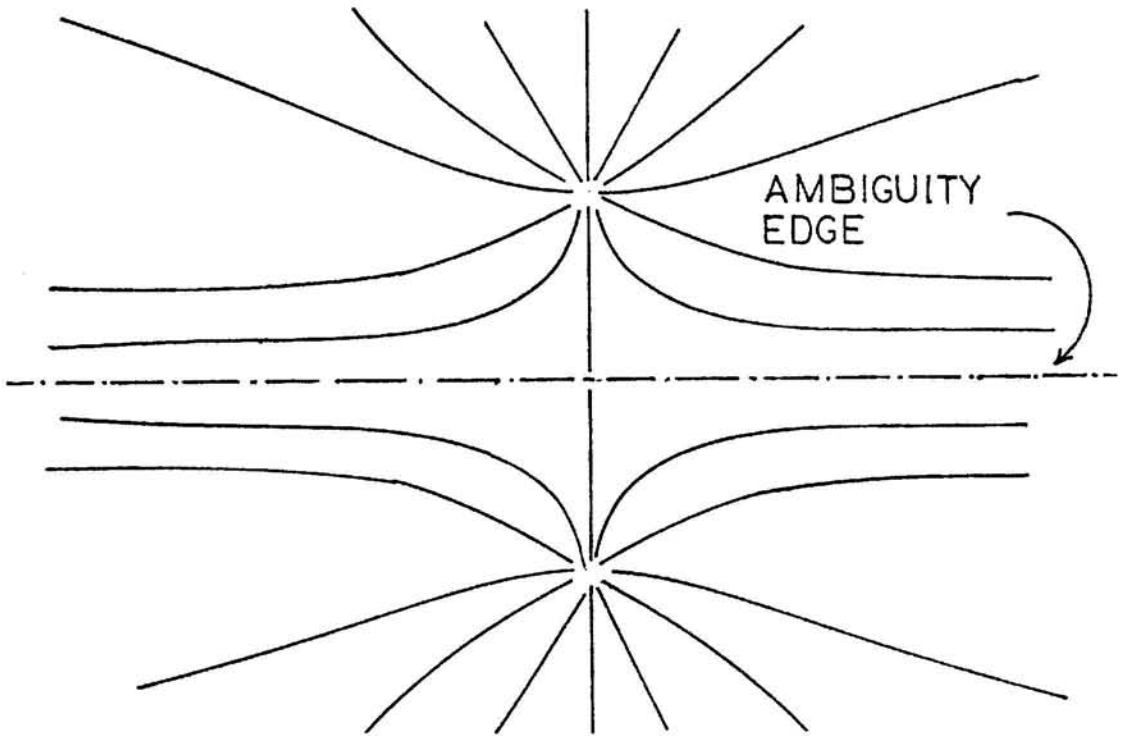


Figure 24: A globally determined ambiguity edge.
$f = 1/(1+x^2 + (y-1)^2) \quad 1/(1+x^2 + (y+1)^2)$

manifests itself in a slowing down of the characteristics as they approach the line $x = 0$ (alternatively $\lambda \to \infty$). They cannot cross this line aggregation of singular points. Note that the characteristics approach this line at right angles and that the edge is determined locally, each point on it being a singular point.

A second kind of ambiguity edge can occur parallel to characteristics, separating those which can be reached from one singular point from those reachable only from another. This kind of edge is not locally determined, since a change in the surface is possible which removes one of the singular points and makes all the characteristics accessible from the other. This can be done without altering an area near two given points previously separated by an ambiguity edge.

Both types of ambiguity edges occur in the general case but are not so easily studied there. They divide the image into regions within each of which a solution can be obtained. Typically most such regions will have one singular point from which one may obtain initial conditions (provided one makes a decision about whether the surface is concave or convex and knows the distance to the singular point).

3.2 LUNAR TOPOGRAPHY:

3.2.1 INTRODUCTION TO LUNAR TOPOGRAPHY:

The other very interesting simplification to the general shape from shading equations occurs when we introduce the special reflectivity function which applies to the material in the maria of the moon. This in fact was the first shape from shading problem solved both theoretically and in an operating algorithm [4]. Using the special reflectivity function and the fact that the sun is a distant source, it is possible (but very tedious) to show that the equations simplify so that the base characteristics (i.e. the projection of the characteristics on the image plane) become straight lines radiating from the zero-phase point. This point corresponds to $g = 0$ and is directly opposite the sun as seen from the camera. Actually this is true only when the sun is located at negative z, for positive z (that is in front of the camera), the relevant point is the phase point, directly in the sun.

3.2.2 REFLECTIVITY FUNCTION FOR THE MARIA OF THE MOON:

The variation of light reflected from the surface of the moon with phase and inclination of the surface has been studied for a long time. At a given lunar phase g, all possible combinations of incident angle i and emittance angle e are represented by some portion of the surface. A fairly good approximation is the Lommel-Seeliger formula [1]:

$$\phi(I,E,G) = \frac{\Gamma_0 \quad (I/E)}{(I/E) + \lambda(G)}$$

Where Γ_0 is a constant and the function $\lambda(G)$ is defined by a table. This formula can also be derived from a simplified model of the lunar surface. A slight gain in accuracy is possible if Γ_0 is allowed to vary with G as well. In particular Fesenkov [1] finds the more accurate formula:

$$\phi(I,E,G) = \frac{\Gamma_0 * (I/E)(1 + \cos^2(\alpha/2))}{(I/E) + \lambda_0(1 + \tan^2(\alpha/2))}$$

Where as before:

$$\tan(\alpha) = \frac{G - (I/E)}{\sqrt{1 - G^2}}$$

A recent theoretical model is that of Hapke [3] which corresponds fairly closely to the measured reflectivity

function. In most of these formulaes we find that for a given G, ϕ is constant for constant I/E. The lines of constant I/E are meridians.

At full moon, when G = 1 we find that the whole face has constant luminosity. This is quite unlike the effect on a sphere coated with a typical matt paint where the image intensity would vary as:

$$\sqrt{1 - (r/R)^2}$$

Where R is the radius of the image and r the distance from the centre of the image. The full moon thus has the same appearance as a flat disc if one is used to objects with normal matt surfaces. This may explain the flat appearance of the full moon.

3.2.3 DERIVATION OF THE SOLUTION FOR LUNAR TOPOGRAPHY:

3.2.3.1 THE BASE CHARACTERISTICS:

In the case of pictures taken of the lunar surface from nearby (e.g. from orbit) we have the following:

1. Distant source (the moon subtends an angle of about .03 milli-radians at the sun).

2. Near point source (the sun subtends an angle of about 10 milli-radians at the moon).

3. Camera at the origin.

4. The reflectivity function is constant for constant I/E. This is a property of the material of the maria of the moon which has been known for some time.

We have (using results obtained in subsection 2.3.4):

$$I_{\underset{\sim}{r}} = 0 \qquad I_{\underset{\sim}{n}} = (1/n)(\hat{\underset{\sim}{r}}_o - I\,\hat{\underset{\sim}{n}})$$

$$E_{\underset{\sim}{r}} = (1/r)(\hat{\underset{\sim}{n}} - E\,\hat{\underset{\sim}{r}}) \qquad E_{\underset{\sim}{n}} = (1/n)(\hat{\underset{\sim}{r}}_o - E\,\hat{\underset{\sim}{n}})$$

$$G_{\underset{\sim}{r}} = (1/r)(\hat{\underset{\sim}{r}}_o - G\,\hat{\underset{\sim}{r}}) \qquad G_{\underset{\sim}{n}} = 0$$

Where $\hat{\underset{\sim}{r}}_o$ is a unit vector in the direction from the sun to the moon.

If I and E depend on some parameter s, while I/E is constant:

$$EI_s = IE_s$$

Since $\phi(I,E,G)$ is constant for constant I/E:

$$\phi_I I_s + \phi_E E_s = 0 \quad \text{and therefore:}$$

$$I\,\phi_I + E\,\phi_E = 0$$

If I and E depend on some parameter k:

$$\phi_I I_k + \phi_E E_k = \phi_I(I_k - (I/E)E_k) = (\phi_I/E)(EI_k - IE_k)$$

$$\phi_{I/E} = \phi_I E = -(E^2/I)\,\phi_E$$

$$\phi_I I_k + \phi_E E_k = E^2\phi_{I/E}(EI_k - IE_k)$$

Using some of our previous results we find:

$$E\underset{\sim}{I}_r - I\underset{\sim}{E}_r = -(I/r)(\hat{\underset{\sim}{n}} - E\,\hat{\underset{\sim}{r}})$$

$$E\underset{\sim}{I}_n - I\underset{\sim}{E}_n = (E/n)\hat{\underset{\sim}{r}}_o - (EI/n)\hat{\underset{\sim}{n}} - (I/n)\hat{\underset{\sim}{r}} + (EI/n)\hat{\underset{\sim}{n}}$$

$$= (1/n)[(\hat{\underset{\sim}{r}}\cdot\hat{\underset{\sim}{n}})\hat{\underset{\sim}{r}}_o - (\hat{\underset{\sim}{r}}_o\cdot\hat{\underset{\sim}{n}})\hat{\underset{\sim}{r}}] = (1/n)(\hat{\underset{\sim}{r}}\times\hat{\underset{\sim}{r}}_o)\times\hat{\underset{\sim}{n}}$$

$$= \frac{1}{n^2 rr_o}\,(\underset{\sim}{r}\times\underset{\sim}{r}_o)\times\underset{\sim}{n}$$

And since $A\underset{\sim}{I}_r = 0$:

$$F\underset{\sim}{I}_r = A\,\phi\underset{\sim}{I}_r - b\underset{\sim}{I}_r$$

$$F\underset{\sim}{I}_n = A\,\phi\underset{\sim}{I}_n$$

We will ignore $F\underset{\sim}{I}_r$ for now, mainly because it has an ugly looking expansion.

$$\underset{\sim}{F_n} = AE^2 \oint_{I/E} \frac{1}{n^2 r r_o} (\underset{\sim}{r} \times \underset{\sim}{r_o}) \times \underset{\sim}{n}$$

$$(\underset{\sim}{r} \times \underset{\sim}{r_o}) \times \underset{\sim}{n} =$$

$$[x_o(z-qy)-x(z_o-qy_o), y_o(z-px)-y(z_o-px_o), z(x_o p+y_o q)-z_o(xp+yq)]$$

Let $X = -AE^2 \oint_{I/E} \frac{zz_o}{n^2 r r_o}$ where $(x_o, y_o, z_o) = \hat{\underset{\sim}{r}}_o$.

Note that A is a constant in this case.

$$F_p = X\left(\frac{x_o}{z_o}(1 - q\frac{y}{z}) - \frac{x}{z}(1 - q\frac{y_o}{z})\right)$$

$$F_q = X\left(\frac{y_o}{z_o}(1 - p\frac{x}{z}) - \frac{y}{z}(1 - p\frac{x_o}{z})\right)$$

Now looking back at the five O.D.E.'s:

$$\dot{x} = F_p, \quad \dot{y} = F_q, \quad \dot{z} = pF_p + qF_q = p\dot{x} + q\dot{y}$$

$$\dot{p} = -F_x - pF_z, \quad \dot{q} = -F_y - qF_z$$

Again we can decide to ignore $\dot{p}$ and $\dot{q}$ for the time being, and attempt to determine the behavior of the characteristics. Our aim is to show that their projections in the image plane are straight lines independent of the scene. The behavior of $\dot{y}$ against $\dot{x}$ is of little help and we next look at the projections in the image plane:

$$x' = x\ (f/z) \quad \text{and} \quad y' = y\ (f/z)$$

$$(\dot{x}'/f) = (1/z^2)(\dot{x}z - x\dot{z}) = (1/z^2)[\dot{x}z - x(p\dot{x} + q\dot{y})]$$

$$= \frac{1}{z}\left((1 - \frac{x}{z}p)\dot{x} - \frac{x}{z}q\dot{y}\right) -$$

$$= \frac{X}{z}\left(\frac{x_o}{z_o}(1 - p\frac{x}{z} - q\frac{y}{z}) - \frac{x}{z}(1 - p\frac{x}{z} - q\frac{y}{z})\right)$$

$$\frac{\dot{x}'}{f} = \frac{X}{z}\left((\frac{x_o}{z_o} - \frac{x}{z})(1 - p\frac{x}{z} - q\frac{y}{z})\right)$$

Similarly:

$$\frac{\dot{y}'}{f} = \frac{X}{z}\left((\frac{y_o}{z_o} - \frac{y}{z})(1 - p\frac{x}{z} - q\frac{y}{z})\right)$$

Now if the surface is not tangent to the ray from the camera:

$$E \neq 0 \quad \text{i.e.} \quad \underset{\sim}{r}.\underset{\sim}{n} \neq 0 \quad \text{and therefore:}$$

$$(1 - p\frac{x}{z} - q\frac{y}{z}) \neq 0$$

If in addition $\phi_{I/E} \neq 0$, $A \neq 0$ and $z \neq 0$, then we can divide the two equations:

$$\frac{\dot{y}'}{\dot{x}'} = \frac{\frac{y_o}{z_o} - \frac{y}{z}}{\frac{x_o}{z_o} - \frac{x}{z}}$$

This first-order ordinary differential equation for the base characteristics is separable:

$$\frac{-dy'}{\frac{y_o}{z_o} - \frac{y'}{f}} = \frac{-dx'}{\frac{x_o}{z_o} - \frac{x'}{f}}$$

Solving this we obtain:

$$\log\left(\frac{y_o}{z_o} - \frac{y'}{f}\right) = \log\left(\frac{x_o}{z_o} - \frac{x'}{f}\right) + \log(c)$$

Let the arbitrary constant c be tan(t):

$$\frac{1}{\sin(t)}\left(\frac{y_o}{z_o} - \frac{y'}{f}\right) = \frac{1}{\cos(t)}\left(\frac{x_o}{z_o} - \frac{x'}{f}\right)$$

Thus the projections of the characteristics are straight lines in the image plane emanating from the point:

$$\left(\frac{x_o}{z_o}, \frac{y_o}{z_o}\right)$$

If the sun is behind the plane of the image ($z_o > 0$ - as would usually be the case for reasonable illumination and avoidance of extraneous light entering the lens) this point is called the zero-phase point, since it corresponds to the point in the scene which is directly opposite the sun as seen from the camera and hence $g = 0$. Because of the special properties of the reflectivity function of the maria of the moon intensity variations in this region are entirely due to

non-uniform surface properties rather than shape. It is for this reason that this point is not usually included in the image but lies somewhat outside it in the image-plane. This will prove unfortunate later on when we have to invent initial conditions.

If the sun is in front of the image plane ($z_0 < 0$), the special point is the TT phase point, where the image of the sun would appear in the image-plane.

So the obvious simplification to the equations which would arise if we let $x_0 = y_0 = 0$ cannot be exploited since we do not wish to orient the camera in this way.

We would like to arrange for s, the parameter that varies along each characteristic, to correspond to arc-length. This can be achieved by multiplying each of the five O.D.E.'s by λ, where:

$$\lambda = -\frac{z}{X} \times \frac{1}{s} \times \frac{1}{\left(1 - p\frac{x}{z} - q\frac{y}{z}\right)}$$

$$\frac{x'}{f} = -\left(\frac{x_0}{z_0} - \frac{x}{z}\right)\frac{1}{s} = -\left(\frac{x_0}{z_0} - \frac{x'}{f}\right)\frac{1}{s}$$

Then by choosing constants suitably:

$$\frac{x}{z} = \frac{x'}{f} = \frac{x_o}{z_o} + s\cos(t)$$

$$\frac{y}{z} = \frac{y'}{f} = \frac{y_o}{z_o} + s\sin(t)$$

Thus s gives arc length along the characteristics while the value of t selects a particular characteristic.

3.2.3.2 THE INTEGRAL FOR z:

We next turn to $\dot{z}$ which we would like to find without solving the messy equations for $\dot{p}$ and $\dot{q}$.

$$\dot{z} = p\dot{x} + p\dot{y} = \lambda X \left(p\left(\frac{x_o}{z_o} - \frac{x}{z}\right) + q\left(\frac{y_o}{z_o} - \frac{y}{z}\right) \right)$$

$$= \lambda X s\,(p\cos(t) + q\sin(t))$$

This is a good place to introduce some abbreviations of commonly occuring dot-products:

$$L = \left(\frac{x_o}{z_o}, \frac{y_o}{z_o}\right) \cdot (\cos(t), \sin(t))$$

$$M = (p, q) \cdot \left(\frac{x_o}{z_o}, \frac{y_o}{z_o}\right)$$

$$N = (\cos(t),\ \sin(t)) \ .\ (p,\ q)$$

Note that L is predetermined (i.e. independent of the image) and that L and M tend to 0 if the camera is pointed directly away from the sun (i.e. $x_o = y_o = 0$)

$$(1 - p\frac{x}{z} - q\frac{y}{z}) = (1-M-sN)$$

$$X = \frac{z}{s} \times \frac{-1}{(1-M-sN)}$$

and so: $$\dot{z} = \frac{-zN}{(1-M-sN)}$$

We now attempt to express this in terms of measureable and calculable quantities (s.a. G, I/E, s and t). Since $\phi_{I/E} \neq 0$ and differentiable it must be monotonic and hence have an inverse. That is, given b/A we will be able to calculate I/E (G is known at each point).

$$\underset{\sim}{r}_o = (x_o, y_o, z_o) \quad \text{and} \quad r_o = \sqrt{x_o^2 + y_o^2 + z_o^2}$$

$$\underset{\sim}{r} = (x,y,z) = z\left(\frac{x_o}{z_o} + s\cos(t), \frac{y_o}{z_o} + s\sin(t), 1\right)$$

Let $$Q = \sqrt{s^2 + 2sL + \left(\frac{r_o}{z_o}\right)^2}$$

Then $r = z0$

$$\underset{\sim}{n} = (-p, -q, 1) \qquad n = \sqrt{1 + p^2 + q^2}$$

$$\underset{\sim}{n} \cdot \underset{\sim}{r} = z\,(1-M-sN)$$

$$\underset{\sim}{n} \cdot \underset{\sim}{r}_0 = z_0(1-M)$$

Let $$T = sL + \left(\frac{r_0}{z_0}\right)^2$$

$$\underset{\sim}{r}_0 \cdot \underset{\sim}{r} = T\, z\, z_0$$

$$G = \frac{\underset{\sim}{r}_0 \cdot \underset{\sim}{r}}{r_0 r} = \frac{z z_0}{r r_0} T = \frac{T\, z_0}{Q\, r_0}$$

So we can calculate G for each point on the characteristic, independent of t and the scene. Next we attempt to rewrite the expression for z in terms of I/E:

$$I/E = \frac{\underset{\sim}{n} \cdot \underset{\sim}{r}_0}{\underset{\sim}{n} \cdot \underset{\sim}{r}} \times \frac{r}{r_0} = \frac{z_0(1-M)}{z(1-M-sN)} \times \frac{zQ}{r_0}$$

$$= Q \frac{z_0}{r_0}\left(1 + \frac{sN}{(1-M-sN)}\right)$$

$$\dot{z} = -\frac{z}{s}\left(\frac{(I/E)}{Q} \times \frac{r_0}{z_0} - 1\right)$$

As mentioned before one can find an inverse Ψ to ϕ s.t.:

$$\Psi(b/A,G) = I/E$$

$$\dot{z} = -\frac{z}{s}\left(\Psi\left(\frac{b}{A},G\right)*\left(\frac{r_o}{z_o}\right)\frac{1}{Q} - 1\right)$$

The usual tables for ϕ in the case of the maria of the moon however are not usually given in terms of I/E and G, but rather α and g. Where :

$$\tan(\alpha) = \frac{G-I/E}{\sqrt{1-G^2}}$$

α is the projection of the emittance angle on the phase-angle plane.

$$G - I/E = \frac{z_o}{r_o}x\frac{T}{Q} - Q\frac{z_o}{r_o}x\frac{(1-M)}{(1-M-sN)}$$

$$= \frac{z_o}{r_o}\frac{1}{Q}\left(\frac{r_o}{z_o} + sL - \frac{(1-M)}{(1-M-sN)}Q^2\right)$$

$$= -\frac{z_o}{r_o}x\frac{s}{Q}\left((s+L) + \frac{N}{(1-M-sN)}Q^2\right)$$

$$1-G^2 = 1 - \frac{z_o}{r_o}x\frac{T^2}{Q^2}$$

Define $$P = \mathrm{sgn}(z_o)\sqrt{\left(\frac{r_o}{z_o}\right)^2 - L^2}$$

$$1-G^2 = \left(\frac{s\ z_o}{Q\ r_o} P\right)^2$$

$$\frac{G-I/E}{\sqrt{1-G^2}} = \frac{1}{P}\left(\frac{N}{(1-M-sN)}\ Q^2 + (s+L)\right)$$

3.2.3.3 THE INTEGRAL FOR r:

So far we have been working in the coordinates x', y' and z. The final result looks neater if we use r instead of z.

$$r = zQ$$

$$\dot{r} = \dot{z}Q + z(s+L)/Q$$

$$= \frac{zN}{(1-M-sN)}\ Q + z\ \frac{(s+L)}{Q}$$

$$= \frac{r}{Q^2}\left(\frac{N}{(1-M-sN)}\ Q^2 + (s+L)\right)$$

$$\dot{r} = -(rP/Q^2)\ \tan(\alpha)$$

Written out more fully, we have:

$$\frac{\dot{r}}{r} = \frac{\mathrm{sgn}(z_o)\sqrt{(\frac{r_o}{z_o})^2 - L^2}}{s + 2sL + (\frac{r_o}{z_o})^2} \tan(\alpha)$$

The numerator is a fixed quantity for each characteristic, the denominator varies along each characteristic (but is independent of the scene), while $\tan(\alpha)$ is obtained from the measurement of b/A and the known G (using the function Ψ). The given ordinary differential equation for r has the simple solution:

$$\frac{r(s)}{r(0)} = e^{-P\int_0^s \frac{\tan(\alpha)}{Q^2}\, ds}$$

where $$L = \frac{x_o}{z_o}\cos(t) + \frac{y_o}{z_o}\sin(t)$$

and $$Q = \sqrt{s^2 + 2sL + (\frac{r_o}{z_o})^2}, \quad P = \mathrm{sgn}(z_o)\sqrt{(\frac{r_o}{z_o})^2 - L^2}$$

$r(0)$ is the distance to the point from where the integration was started.

To sum up: as one advances along each characteristic in turn, one calculates G, measures b/A and uses Ψ to obtain

tan(α), which is then used in the evaluation of the above integral. The process is much simpler than the general shape from shading algorithm in that the base characteristics are predetermined straight lines and only an integral needs to be evaluated. It is possible to write the above result in a slightly more elegant form, which is the one derived by T. Rindfleisch (for $z_0 > 0$):

$$\frac{r(P)}{r(P_0)} = e^{-\frac{1}{f|\hat{z}\times\hat{N}_0|}\int_0^S (\hat{r}'\cdot\hat{z})\ \tan(\alpha)\ ds'}$$

Where $s' = fs$

$$\underset{\sim}{N}_0 = \underset{\sim}{r}_0\times\underset{\sim}{r}' = -(f/z)(\underset{\sim}{r}\times\underset{\sim}{r}_0)$$

$$= \frac{f}{z} z z_0 \left(\left(\frac{y}{z} - \frac{y_0}{z_0}\right), \left(\frac{x_0}{z_0} - \frac{x}{z}\right), \left(\frac{xy_0}{zz_0} - \frac{x_0 y}{z_0 z}\right) \right)$$

$$= -f z_0 s\left(-\sin(t), \cos(t), \frac{y_0}{z_0}\cos(t) - \frac{x_0}{z_0}\sin(t)\right)$$

Now $$L^2 = \frac{x_0^2}{z_0^2}\cos^2(t) + 2\frac{x_0 y_0}{z_0 z_0}\cos(t)\sin(t) + \frac{y_0^2}{z_0^2}\sin^2(t)$$

$$N_0 = f z_0 s\sqrt{1 + \frac{y_0^2}{z_0^2}\cos(t) - 2\frac{x_0 y_0}{z_0 z_0}\cos(t)\sin(t) + \frac{x_0^2}{z_0^2}\sin(t)}$$

$$= f z_0 s P$$

$$\hat{\underset{\sim}{z}} \times \underset{\sim}{N}_o = -f z_o s\,(\cos(t),\ \sin(t),\ 0)$$

$$|\hat{\underset{\sim}{z}} \times \underset{\sim}{N}_o| = s f |z_o|$$

$$\frac{1}{|\hat{\underset{\sim}{z}} \times \hat{\underset{\sim}{N}}_o|} = \frac{N_o}{|\hat{\underset{\sim}{z}} \times \underset{\sim}{N}_o|} = P$$

$$(\hat{\underset{\sim}{r}}' \cdot \hat{\underset{\sim}{z}})^2 = (\hat{\underset{\sim}{r}} \cdot \hat{\underset{\sim}{z}})^2 = (\underset{\sim}{r} \cdot \hat{\underset{\sim}{z}})^2 / r^2 = (z/r)^2 = 1/Q^2$$

The two ways of writing the integral are thus equivalent.

3.2.4 SOME COMMENTS ON THE INTEGRAL SOLUTION:

1. The base characteristics are predetermined straight lines (independent of the image). This makes for high accuracy and ease in planning a picture taking mission.

2. Only a single integral needs to be evaluated, not five differential equations.

3. The primary input is the intensity, not its gradients, again making for high accuracy.

4. Although, as usual, the reflected light-intensity does not give a unique normal, it does determine the slope component in the direction of the characteristic.

J. van Diggelen [2] first noted a special case of this when he solved the lunar topography problem for the special case of an area near the terminator (line separating sunlit from dark areas). The characteristics are such that the slope along them can be determined locally . The slope at right angles to the characteristics cannot be determined locally.

5. Although T. Rindfleisch [4] did not mention it in his paper it is very easy to bridge shadows since each light-ray lies in a sun-camera-characteristic plane. Its image can thus be traced on the base characteristic until we again meet a lighted area. One need not even make special provisions for this, but just use $\tan(\alpha)$ for grazing incidence (intensity = 0) in the shaded section.

3.3 APPLICATION TO OBJECTS BOUNDED BY PLANE SURFACES:

Since a great deal of image processing these days is applied to images of polyhedra one might enquire how one could apply this method to such objects. First we note that the main features of these objects, the joints (angular edges on an object) and edges (where one object occludes another), are a stumbling block to the application of our method developed so

far. Since we already know that the areas of more or less constant reflectivity are plane faces there is little point in exploring them. So a completely different approach is indicated. Firstly we might check whether a parsing of the scene obtained by some other method is correct in the sense that the intensities observed for the faces correspond to their inclinations (the information on the intensity of the faces is usally discarded). It is not clear however what one might do if this test fails. Furthermore, the programs which reduce the image to a line-drawing and then decide which faces belong to each object cannot really determine the inclinations of the various faces without additional assumptions (orthogonality for example).

One can however find the normals to each face directly from the known slopes of the projection of the joints in the image and the measured reflectivities. One must know which lines in the image are true joints (between two faces belonging to the same polyhedron) and which are fortuitous (edges between faces of different objects). For each normal we need two values. Each intensity gives us one non-linear equation and each slope of an image of a joint gives us another. The equation from the intensity is of course:

$$A(\underset{\sim}{r})\ \phi(I,E,G) - b(\underset{\sim}{r}') = 0$$

Where we know that I, E and G are functions of p and q. There will be one such equation for each face.

Where two faces with normals $\underset{\sim}{n}_1$ and $\underset{\sim}{n}_2$ intersect, they form a joint which will be seen in the image. Suppose two points on this image are $\underset{\sim}{A}$ and $\underset{\sim}{B}$. Then a vector perpendicular to the plane through the joint and the camera is $\underset{\sim}{A} \times \underset{\sim}{B}$. We also know that the joint must be parallel to $\underset{\sim}{n}_1 \times \underset{\sim}{n}_2$. But $\underset{\sim}{A} \times \underset{\sim}{B}$ must be perpendicular to the joint hence:

$$(\underset{\sim}{A} \times \underset{\sim}{B}) \cdot (\underset{\sim}{n}_1 \times \underset{\sim}{n}_2) = 0$$

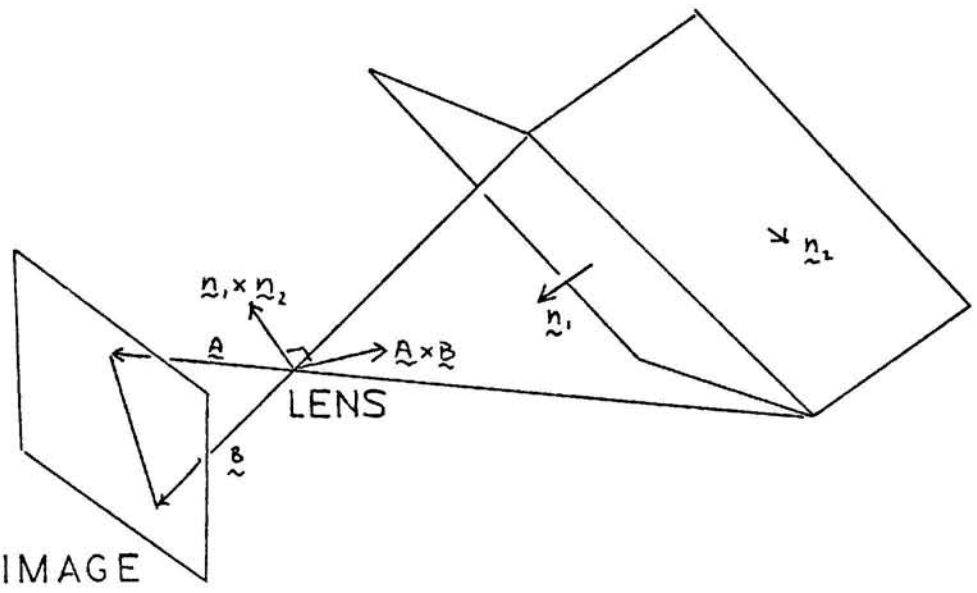


Figure 25: Projection of a joint on a polyhedron on the image.

Each joint contributes one such equation. Next we determine how many faces must intersect before we have enough information for a solution. Two faces intersecting give us two equations from the intensities and one from the slope of the image of the joint, while we need four unknowns. An infinity of solutions thus exist. With three faces a solution is possible since we have six equations in six unknowns. Because of the non-linearity of the equations more than one solution might exist. With a larger number of faces we always have at least enough information for a solution and at times have more equations than unknowns which may remove some of the remaining ambiguities and improve the accuracy. In this way too it may be possible to discover which joints are really between faces of the same object.

3.4 FACIAL MAKE-UP:

When a surface whose photometric properties are taken to be uniform is treated so as to change these properties in some areas, the apparent shape is changed. This of course is one of the uses of make-up. The shape of a face for example can be made to conform more closely to what a person thinks is currently considered 'ideal'. This is achieved by making some areas darker (causing them to appear steeper) and others

Figure 26: Illustration of the effect of facial make-up.

Figure 27: Illustration of the effect of facial make-up.

lighter. Areas lightened usually include singular points and cause a change in the apparent skin darkness (a normalization effect) and will change the apparant shape in areas other than the singular points.

These modifications can change the shape perceived when viewed under the right lighting conditions. The effect will change somewhat with orientation and may at times disappear when no reasonable shape would give rise to the shading observed. Because of a number of surface oils the skin has a specular component in its reflectivity, it is also fairly translucent. Both of these effects are sometimes controlled with talcum powder. The removal of the specular components makes the surface appear more smooth.

4. EXPERIMENTAL RESULTS:

4.1 A PROGRAM SOLVING THE CHARACTERISTICS SEQUENTIALLY:

When the solution to the shape from shading problem had been found using the inconvenient coordinate system (x', y', z), a program was written which would solve the five O.D.E.'s along one characteristic at a time (the equations used are not reproduced here). The input data was obtained from the image-dissector camera attached (at that time) to the PDP-6 computer in the Artificial Intelligence Laboratory. This camera is a random access device: when given an x and a y coordinate it returns a number proportional to the intensity at that point in the image. The program first searches for a maximum in intensity, constructs a small spherical cap around it (to obtain an initial curve) and uses a standard numerical method (see subsection 4.1.2) to solve the set of five ordinary differential equations.

The prime data required in this solution is the intensity gradient (x' and y' derivatives of the intensity), which is obtained from the intensities measured for a small raster of points near the current x' and y'. A linear function in x' and y' is fitted to this set of intensities; the coefficients of x' and y' are the desired gradients. The

size of the raster is chosen to correspond to the step-size used in the numerical solution method, so that successive rasters almost, but not quite, touch. In this way fair accuracy in the determination of the gradients is obtained without loss in resolution.

If the least-squares fit is bad, indicating that surface detail is being missed with the stepsize used, or that the characteristic is traversing an edge or joint, the solution for this characteristic is halted and the solution started for the next characteristic. Other reasons for terminating the solution are that the characteristic has left the field of view of the image-dissector or reached a very dark region, most likely a shadow or the background. When either the (calculated) incident or emittance angles become very small (indicating approach to an edge or shadow edge) or λ very large (indicating approach to another singular point or an ambiguity edge) the solution will also be stopped.

The data structure here is very simple; just a record of various values (x', y', z, intensity, p' and q') for each point on each characteristic. The shape so determined can be displayed in perspective and stereo on a DEC *340* display. The characteristics appear as dashed lines - each dash representing a step in the integration (We chose the

parameter s so that each dash represents the same arc-length). The output can be photographed from the display and plotted on a Calcomp plotter.

4.1.1 AUXILIARY ROUTINES:

A number of auxiliary routines needed to be written for this program. First the Incompatable Time Sharing System (ITS) on the PDP-6 does not support a FORTRAN style arithmetic language and all programming was done in assembly language (MIDAS). The large amount of arithmetic involved, particularly with the inconvenient notation and coordinate system used at first, made it imperative to incorporate into the assembler the ability to handle arithmetic statements.

Next we constructed a package of useful routines which handles floating point I/O, dynamic array allocation and easy generation of display lists for the DEC 340. It also includes routines for the standard arithmetic functions (SQRT, SIN, LOG etc.) and manipulation of vectors and matrices (multiplication, addition, inversion etc.). Interrupts, user defined operations and command interpretation are dealt with as well. Some of the remaining routines will be briefly described in the next few sections.

4.1.1.1 STEREO PROJECTION AND OBJECT ROTATION:

Since it is important (particularly during the debugging phase) to be able to visualize the shape being calculated, stereoscopic output on the display is provided.

Let d_s be the separation of the eyes, f their distance from the display and d_o the distance from the eyes to the origin of the coordinate system (usually chosen to be at the singular point from which the solution was started). The coordinates of the left-eye and right-eye images of the point (x,y,z) are then (x'_l,y') and (x'_r,y') where:

$$x'_{l,r} = (x \pm d_s/2)(f/(z+d_o)) \mp d_s/2$$
$$y' = y\,(f/(z+d_o))$$

A pair of lenses is employed to focus on the surface of the display while converging on the apparent point (x,y,z). Obviously one needs to know the scaling of the display in terms of dsiplay units per mm.

One would like to be able to view the objects from various sides and perhaps even have some rotational motion to gain a greater perception of depth. To this end the object can be rotated around its origin (also offset and expanded in size).

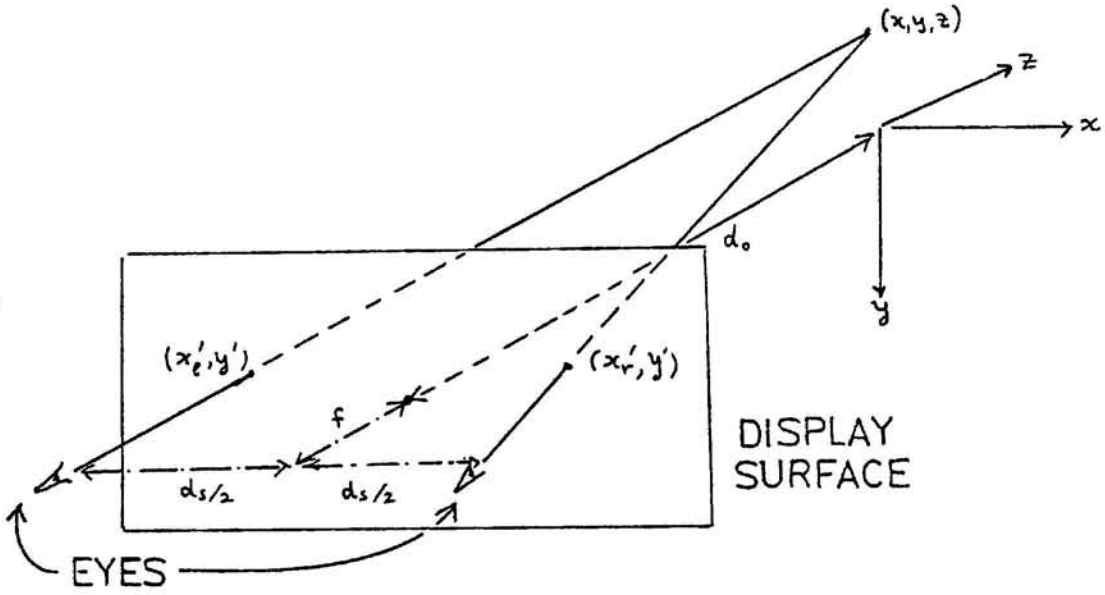


Figure 28: Stereo-projection of an object point.

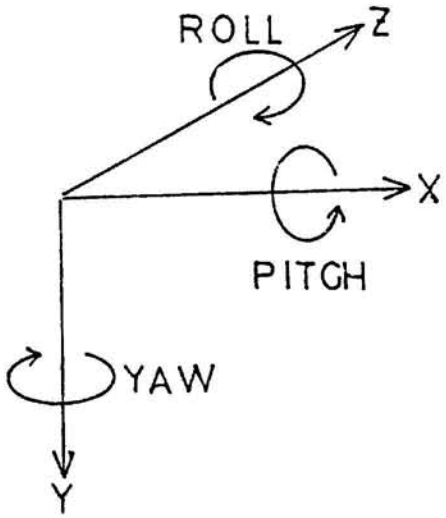


Figure 29: Definition of Pitch, Yaw and Roll.

This was preferred over the more common method of allowing the eyes to be moved around in the object space.

To obtain any orientation with one matrix multiplication, the three angles P (pitch), Y (yaw) and R (roll) are defined as rotations about the x, y and z axes respectively. They are applied in that order (order is important because rotations are not commutative).

$$A = \begin{pmatrix} \cos(R) & -\sin(R) & 0 \\ \sin(R) & \cos(R) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(Y) & 0 & \sin(Y) \\ 0 & 1 & 0 \\ -\sin(Y) & 0 & \cos(Y) \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(P) & -\sin(P) \\ 0 & \sin(P) & \cos(P) \end{pmatrix}$$

Using the abbreviation c for cosine and s for sine we have:

$$A = \begin{pmatrix} cR\,cY & cR\,sY\,sP - sP\,cP & cR\,sY\,cP + sR\,sP \\ sR\,cY & sR\,sY\,sP + cR\,cP & sR\,sY\,cP - cR\,sP \\ -sY & cY\,sP & cY\,cP \end{pmatrix}$$

The various parameters controlling the object rotation and the projective mapping can either be preset or continously read in from a number of potentiometers (connected to a multiplexor and an A/D convertor) controlled by the viewer. While one display list appears, the other is being calculated using the latest set of parameters and will in turn be displayed when completed. The parameters are also displayed, they are:

PITCH, YAW and ROLL (P, Y and R)

SIZEC or MAG - magnification of the object

FDIS or DIMG - distance from eye to display (f)

DOBJ - distance from eye to object (d_o)

DSEY or EYES - seperation of the eyes (d_s)

For photographic purposes each of the two images in turn can be blown up (to account for the reduction in size in the camera) and displayed a fixed number of times while the shutter is open. Windowing at the edge of the screen is automatic and some very simple kinds of hidden line elimination are available but not normally used. The same stereo display package is used by the later version of the program (new SHADE).

4.1.1.2 MEASURING THE REFLECTIVITY FUNCTION:

The reflectivity functions of some paints were measured using spheres (large rubber balls) as calibration objects. Both camera and source were moved as far away as possible to achieve almost constant phase angle g. The image of a convex object is especially useful because it contains two points for all possible combinations of the incident and emittance angles (i and e) for a given phase angle (g). The position

of the light-source is measured, as well as the distance from the front of the sphere to the entrance pupil. The image-dissector is focused on the edge of the sphere.

With the sphere temporarily illuminated from several sources, the program finds its exact position and size, as well as the difference in horizontal and vertical deflection sensitivity of the image-dissector. It is now in a position to calculate the points in the image which correspond to given incident and emittance angles. For a number of choices of both of these angles it then reads the intensity at a small raster of points (to reduce noise and the effect of pin-holes in the photo-cathode) near these positions and averages them. Since there are usually two places in the image with the same incident and emittance angle, a check on the data is available. The resultant table of values (usually normalized w.r.t. the brightest intensity) can be printed and the whole process repeated after moving the light-source to a new position for a new phase angle. The program accounts for such things as change in incident light intensity as the light-source gets moved around.

4.1.1.3 FINDING THE CALIBRATION SPHERE:

This subsection and the next deal with details, needed in the program for measuring the reflectivity function, which may not be of general interest.

For good accuracy we first need to know the exit pupil to image plane distance (the focal length is given). This would be easy if one could focus on the front of the sphere. It turns out that a simple approximation will work in a few iterations. At each step one recalculates the estimated distance to the edge of the sphere, the estimated exit pupil to image plane distance and the estimated radius of the sphere, using the previous estimates and the measured radius of the image.

Next we need to find the exact center and radius of the sphere from its image coordinates and the known distance to its front. First consider horizontal coordinates only.

$$x_1 = f \tan(a),\ x_2 = f \tan(a+b) \text{ and } x_3 = f \tan(a+2b)$$

We are given x_1 and x_3 and wish to calculate x_2, which can be done after expanding the tangents.

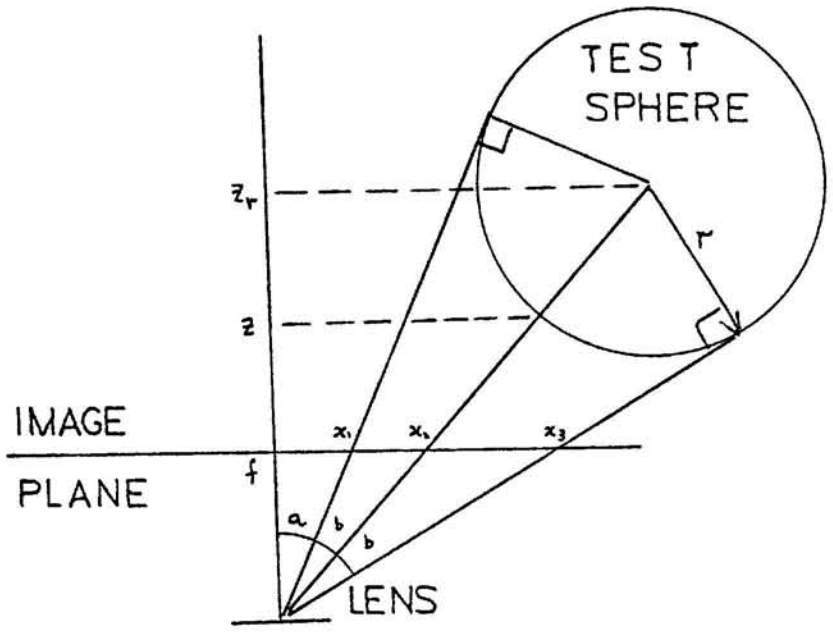


Figure 30: Determining the exact position of the calibration sphere.

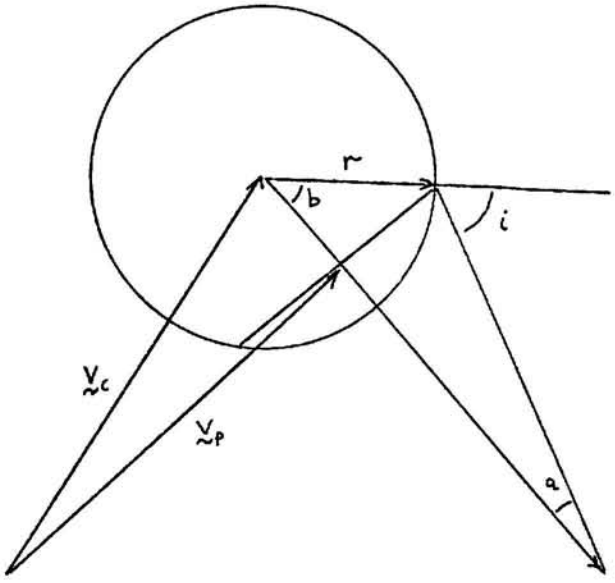


Figure 31: Finding points for given incident and emittance angles.

$$\tan(2b) = (x_3 - x_1)/(f^2 - x_3 x_1)$$

$$\tan(b) = (\sqrt{1 + \tan^2(2b)} - 1)/\tan(2b)$$

$$x_2/f = (x_1 + f\tan(b))/(f - x_1\tan(b))$$

The same formulae are then used to find the vertical position y_2. Finally we need to find z_r:

$$r = R_s \tan(b)\ (\sqrt{1 + \tan^2(b)} + \tan(b))$$

$$z = R_s f/\sqrt{f^2 + x^2 + y^2}$$

$$z_r = z\ (R_s + r)/R_s$$

4.1.1.4 FINDING POINTS FOR GIVEN i AND e:

Clearly the points for given incident angle lie on a circle (on the surface of the sphere). Similarly for points with a given emittance angle. These two circles may intersect in two, one or no points. One can find this intersection by first finding the line along which the planes containing these circles intersect. Applying the sine and cosine laws we get:

Let $I = \cos(i)$ as usual and $D = |\underset{\sim}{v}_s|$

$D/\sin(\pi - i) = r/\sin(a)$ $\quad b = i - a$

$$r\cos(b) = r(\cos(i)\cos(a) + \sin(i)\sin(a))$$

$$= (r/D)\left(I\sqrt{D^2 - r^2(1 - I^2)} + r(1 - I^2)\right)$$

$$d = r\cos(b)$$

$$\underset{\sim}{v}_p = \underset{\sim}{v}_c + d\,\hat{\underset{\sim}{v}}_s$$

The equation of the plane in which the circle of points with given incident angle i lies is:

$$\underset{\sim}{v}\cdot\hat{\underset{\sim}{v}}_s = \underset{\sim}{v}_p\cdot\hat{\underset{\sim}{v}}_s = c \quad \text{say} \qquad (\text{where } \underset{\sim}{v} = (x,y,z))$$

One can find a similar equation for the plane in which the circle of point with given emittance angle e lies. The introduction of an arbitrary third plane allows us to find one point $\underset{\sim}{v}$ on the intersection of the first two. The line of intersection of the first two planes must be parallel to the cross-product of their normals (let them be $\underset{\sim}{v}_{s1}$ and $\underset{\sim}{v}_{s2}$). So the equation of the line we are looking for is:

$$(\underset{\sim}{v} - \underset{\sim}{v}_a) = k\,\underset{\sim}{v}_l \quad \text{where} \quad \underset{\sim}{v}_l = \underset{\sim}{v}_{s1} \times \underset{\sim}{v}_{s2}$$

The points we are trying to find must also lie on the sphere, i.e.:

$$(\underset{\sim}{v} - \underset{\sim}{v}_c)^2 = r^2$$

$$(\underset{\sim}{v}_a + k\,\underset{\sim}{v}_1 - \underset{\sim}{v}_c)^2 = r^2$$

$$k^2 \underset{\sim}{v}_1 . \underset{\sim}{v}_1 + 2\,k\,\underset{\sim}{v}_1 . (\underset{\sim}{v}_a - \underset{\sim}{v}_c) + (\underset{\sim}{v}_a - \underset{\sim}{v}_c) - r^2 = 0$$

The above equation may have no solution for k, in which case no point exists for the given incident and emittance angle. Otherwise we can use the two solutions and substitute back to obtain the desired coordinates which are then transformed into image coordinates.

4.1.1.5 SOME REFLECTIVITY FUNCTIONS:

The first paint investigated was a matt white paint consisting of particles of SiO_2 and TiO_2 suspended in a transparent base. Very roughly one finds that the reflectivity function behaves like cos(i) for a given g. After playing with polynomial fits for a while, the following fairly accurate formula was found by a process of little interest here:

$$\phi(I,E,G) = \frac{(1+G)(2+G)}{6}\left(I + \frac{1+2IEG-(I^2+E^2+G^2)}{16\,(1-G)}\right)$$

Note the appearance of the discriminant discussed in an earlier section (2.1.3). The symbolic manipulation program

	I →									
	1.00	.97	.93	.87	.78	.68	.56	.43	.29	.15
1.00					.77					
.97				.87	.78	.66				
.93			.93	.88	.78	.67	.57			
.87		.97	.94	.89	.79	.67	.57	.45		
E .78	.99	.98	.95	.90	.81	.68	.59	.46	.32	
↓ .68		.98	.95	.91	.82	.71	.59	.47	.33	.18
.56			.94	.90	.83	.74	.61	.48	.34	.17
.43				.88	.79	.74	.62	.50	.34	.18
.29					.79	.70	.58	.42	.30	.15
.15						.65	.50	.38	.26	.13

Figure 32:

Table of reflectivity (for a white matt paint) versus I = cos(i) and E = cos(e) for G = cos(g) = 0.81 . The intervals chosen correspond to constant size steps in the angles. Note the blank areas for combinations of angles which cannot form a spherical triangle (see section 2.1.3)

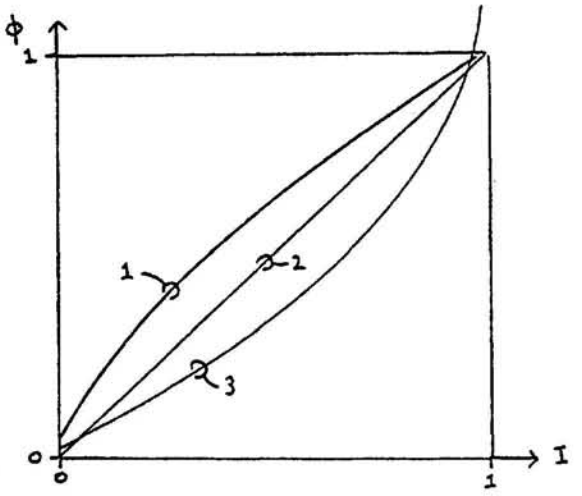
MATHLAB [13] was used to find the derivatives ϕ_I, ϕ_E and ϕ_G needed for the shape-from-shading program. For 'reasonable' angles the above formula is about 5% accurate, becoming worse for extreme angles. The repeatability of this measurement was disappointingly low, depending on the depth of the paint coat and the conditions of its application. Much of the investigation of the behavior of the image-dissector was the result of efforts to trace the remaining causes of inaccuracy.

Some other paints and an eggshell showed a matt component similar to the above, plus a very strong specular component (which is small except near the point for which i = e and i + e = g). This component is very sensitive to small changes in the surface properties such as can be brought about by handling the object.

The image of a convex object with such a surface will usually have two local maxima in intensity. One of these will be broad (corresponding to the matt component), the other narrow and bright (corresponding to the specular component). These may be distinguished by a computer program on the basis of just these properties. It would then be possible to start a solution from the matt maximum (which is not a global maximum) rather than the specular maximum. This might be a

good idea because of the increased accuracy (for one thing the normalization of image intensities would be more accurate).

For the nose-recognition program, a plaster nose was used initially, coated with the matt paint described above. This of course was not suitable for the final experiments. The



1. A matt paint.
2. Lambertian reflec.
3. Skin on nose.

Figure 33: Comparison of some reflectivity functions.

restricted lighting conditions described later were chosen partly to avoid having to find the full-flegded function o of three angles for skin. Since no true sphere covered with skin is available, measurements were taken of the shape of a real nose and intensities in an image (of a transparency) used to estimate $\phi(I,I,1)$. In this way the non-linearities of the photographic process (and they were great) did not have to be determined separately. The properties of skin are

of course not very uniform and also vary from person to person, so no great effort toward accuracy was made. Skin has a highly variable specular component, so any normalization had to be done not w.r.t. the brightest point, but one nearby. The resultant table of $\phi(I)$ versus I lies somewhat below the one obtained from the matt paint under the same lighting conditions.

4.1.1.6 PROPERTIES OF THE IMAGE-DISSECTOR:

In an attempt to track down poor results in the first try at finding reflectivity functions accurately, the image-dissector was investigated in some detail [9]. Amongst problems found were:

1. Unequal deflection sensitivity in horizontal and vertical directions (differed by 12%).

2. Twist of image varying with distance from center of field of view.

3. Poor resolution (3 line-pairs/mm - radius of tube 50 mm).

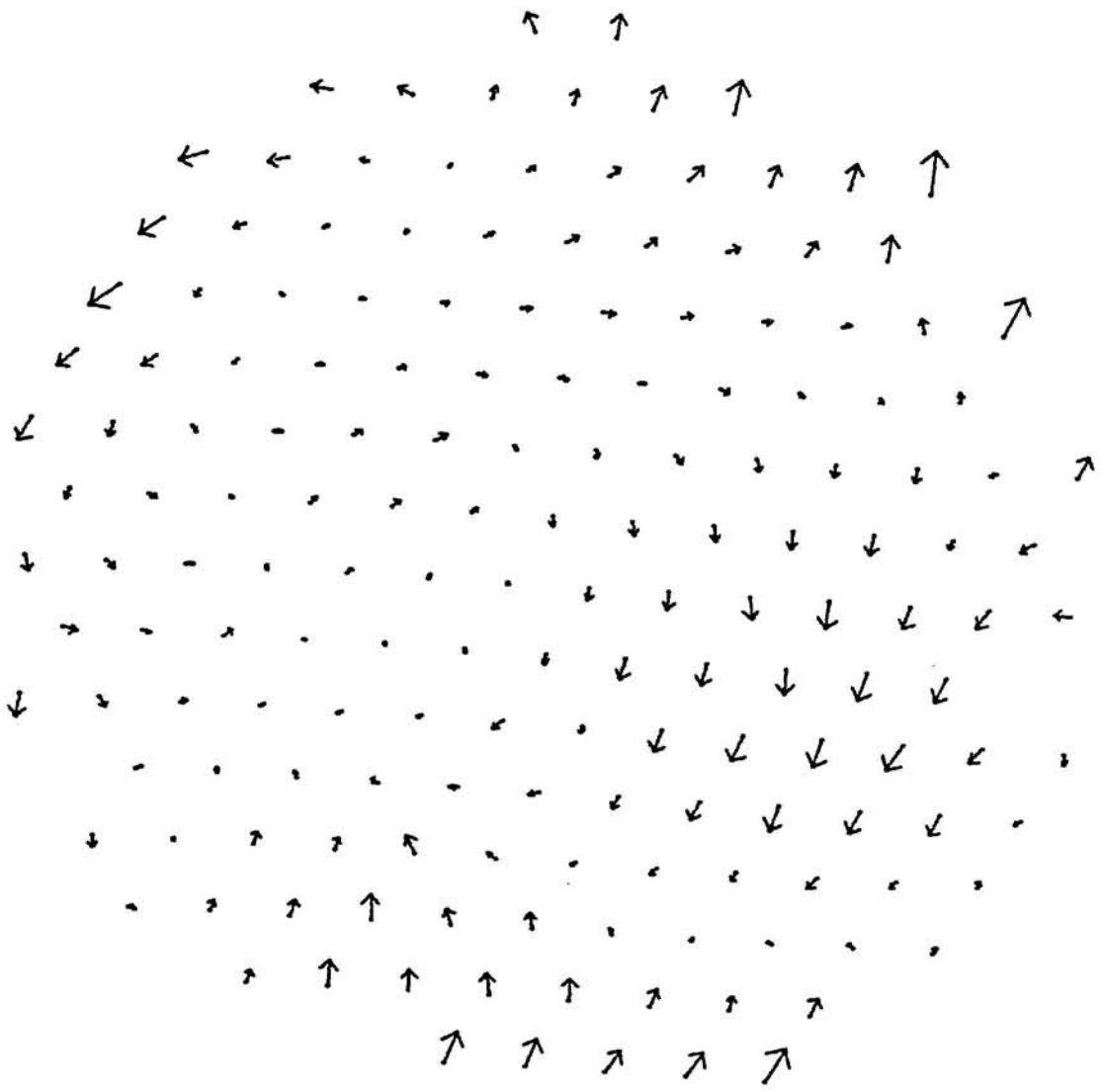


Figure 34: Geometric distortion in image-dissector for a triangular raster of points covering the photocathode. (The arrows are exaggerated 3 times.)

4. Pinholes in the photo-cathode (about 20 of up to 0.5 mm in size).

5. Non-uniform sensitivity of the photo-cathode (varies more than 30%).

6. Fairly long settling time of the deflection coils.

7. A large amount of scatter, which reduces the contrast by almost one-half and causes intensities measured on the image of a uniform square on a dark background to vary by 20%, depending on how close to the edge the measurement is taken.

Some of these difficulties are inherent in the state-of-the-art of these devices, others were repaired. In any case, it was possible now to think about how to improve the program to be more insensitive to these shortcomings.

The program for finding reflectivity functions using spheres as calibration objects was sensitive to the (at that time) severe deflection inaccuracies, since the emittance angle varies rapidly near the edge of the sphere (this effect could be reduced with a parabolic test-object). A calibration table was created by another program in which are recorded

the image-dissector coordinates of a rectangular raster of equally spaced points on the photo-cathode. Also recorded is the sensitivity of the photo-cathode at each point. A simple interpolation routine can then be applied to coordinates sent to the image-dissector to counteract the distortion and, similarly, the intensity values returned can be corrected. A more convenient triangular raster of points covering the whole photo-cathode was later established. Adjustments to the image-dissector eventually reduced the distortion by a significant factor and use of this table was no longer vital, although it did improve accuracy.

4.1.2 NUMERICAL METHODS FOR SOLVING THE O.D.E.'S:

The five O.D.E's were at first solved using a well known Runge-Kutta method [7, page 212]. The idea is that at a given point we can calculate the derivatives of the five variables (x', y', z, p' and q') w.r.t. the parameter s. Using these we take a half-step forward (increment s by $h/2$ and calculate new values for x', y', z, p' and q' as though derivatives higher than the first where zero). We then calculate the derivatives at this new point and take the same step (now using the new derivatives, which will differ slightly from the previous ones). We next take a full step

```
SHADE   SCRIPT  70:05:18  08:15:53
PITCH= 0.400  YAW= 0.000  ROLL= 0.000 SIZEC=  2.0
FDIS= 118.7  DOBJ= 608.2  DSEY=  56.0
```

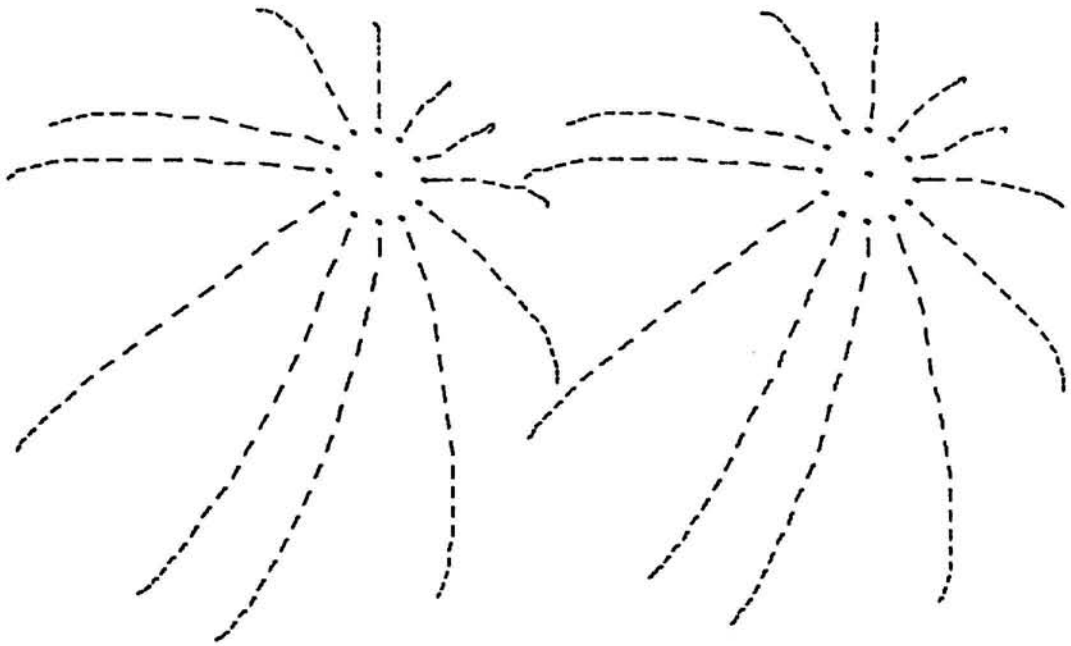


Figure 35: Stereo pair of solution produced by old SHADE.

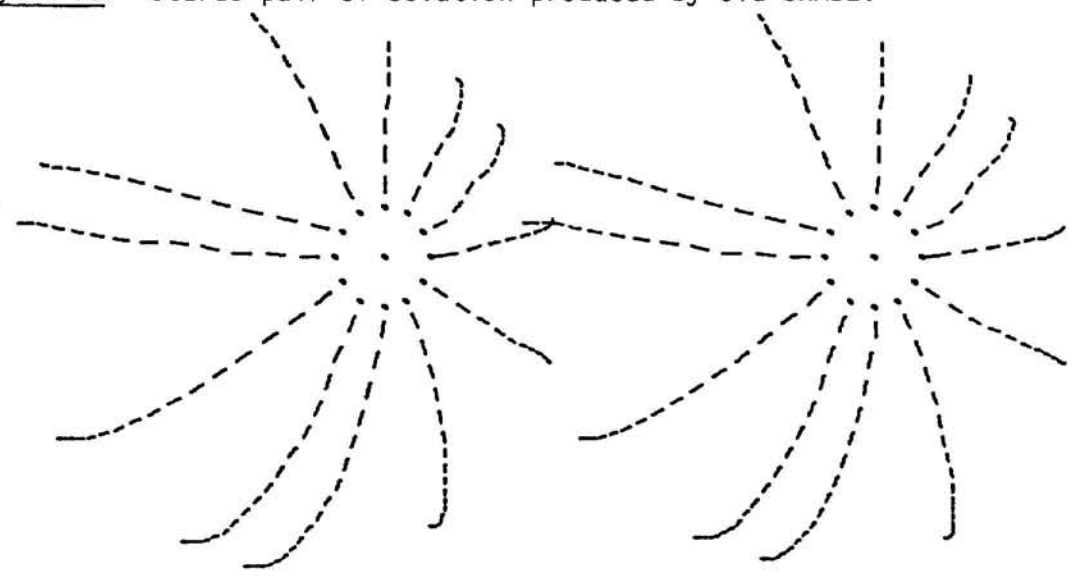


Figure 36: Same solution viewed after a slight rotation.

(increment s by h) The final full step is taken using a weighted average of the four derivatives found in this way. Written out in symbols this becomes:

Let h be the step-size (for the parameter s)

And $\underset{\sim}{Y} = (x', y', z, p', q')$

Also let the equations for the derivatives be:

$$\underset{\sim}{Y}' = F(s, \underset{\sim}{Y})$$

(In our case F is actually independent of s)

Denote $\underset{\sim}{Y}(s_n)$ by $\underset{\sim}{Y}_n$ then at the n^{th} step:

$$\underset{\sim}{K}_1 = h\ F(s_n,\ \underset{\sim}{Y}_n)$$

$$\underset{\sim}{K}_2 = h\ F(s_n + h/2,\ \underset{\sim}{Y}_n + \underset{\sim}{K}_1/2)$$

$$\underset{\sim}{K}_3 = h\ F(s_n + h/2,\ \underset{\sim}{Y}_n + \underset{\sim}{K}_2/2)$$

$$\underset{\sim}{K}_4 = h\ F(s_n + h,\ \underset{\sim}{Y}_n + \underset{\sim}{K}_3)$$

$$\underset{\sim}{Y}_{n+1} = \underset{\sim}{Y}_n + (1/6)(K_1 + 2K_2 + 2K_3 + K_4)$$

This method is easy to start (requires no previous values of $\underset{\sim}{Y}$) and stable, but requires four time-consuming evaluation of the derivatives per step. For this reason various predictor-modifier-corrector methods [7, page 194] were tried and the simplest was found to give adequate accuracy:

$$\underset{\sim}{P}_{n+1} = \underset{\sim}{Y}_n + 2\,h\,F(s_n, \underset{\sim}{Y}_n)$$

$$\underset{\sim}{M}_{n+1} = \underset{\sim}{P}_{n+1} - (4/5)(\underset{\sim}{P}_n - \underset{\sim}{C}_n)$$

$$\underset{\sim}{C}_{n+1} = \underset{\sim}{Y}_n + (h/2)(F(s_n, \underset{\sim}{M}_{n+1}) + F(s_n, \underset{\sim}{Y}_n))$$

$$\underset{\sim}{Y}_{n+1} = \underset{\sim}{C}_{n+1} + (1/5)(\underset{\sim}{P}_{n+1} - \underset{\sim}{C}_{n+1})$$

$\underset{\sim}{P}$, $\underset{\sim}{M}$ and $\underset{\sim}{C}$ are the predictor, modifier and corrector respectively. This method is stable and requires only two derivative evaluations per step, but is not self-starting. The Runge-Kutta method was retained for the first step in the integration. Stability and accuracy were not serious concerns since the noise in the data input contributes far more to errors in the solution.

4.1.3 ACCURACY OBTAINABLE:

Under optimal conditions (using the methods described to cancel out most of the distortion and non-uniformity of photo-cathode sensitivity) the program was allowed to scan a sphere of 100 mm radius. A sphere was then fitted by an iterative least-square method to the data points found. The data points nowhere deviated from the fitted sphere by more than 10 mm, and by less than 5 mm except near the very edge of the image. Such accuracy will not usually be obtained because of non-uniformity in the paint, shortcomings of the

sensing device etc. For many purposes however less accuracy is quite acceptable and for object recognition in particular a more important criterion is most probably that similar objects are distorted in similar ways.

4.1.4 PROBLEMS WITH THE SEQUENTIAL APPROACH:

It soon become apparent that solving the characteristics sequentially had many disadvantages in the general case, even though it works well for lunar topography. The first reason is that as the characteristics spread out from the singular point, they begin to separate and leave large portions of the image unexplored, obtaining only a very uneven sampling of the surface of the object (This is no problem for lunar topography since here the solution is not started from the singular point, but at a place where the spread of the characteristics is small).

With a more parallel approach new characteristics can be interpolated as we go along (and some deleted as they approach too closely).

Next we find that the base characteristics (projections of the characteristics onto the image) may sometimes cross.

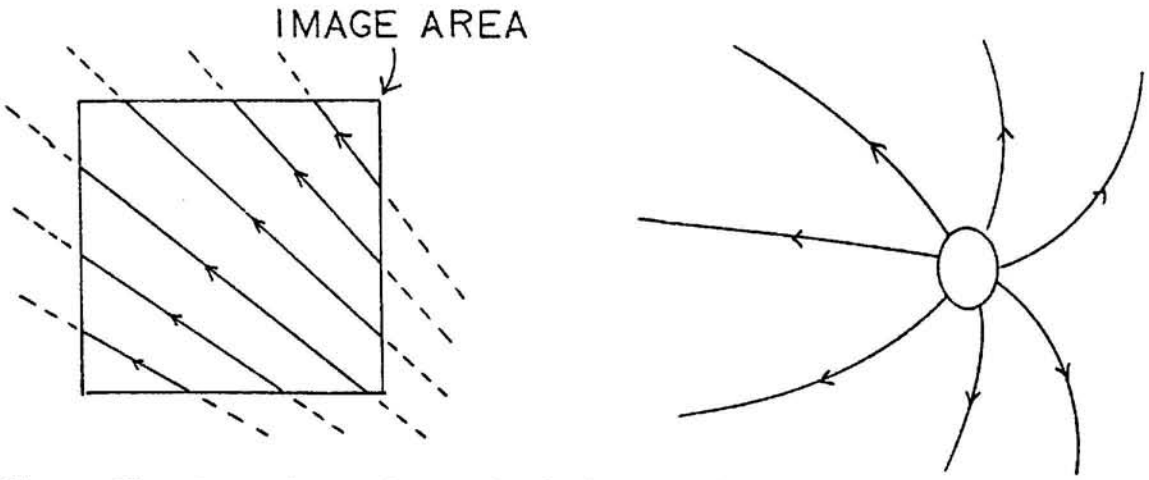


Figure 37: Comparison of spread of characteristics for typical solutions in case of lunar topography and general case.

This is not possible if the solution was exact, since it indicates that the surface is double-valued or at least that its gradient is double-valued. Characteristics may converge or diverge from a (singular) point however. Crossing of characteristics is really symptomatic of another problem which was touched upon when proving the equivalence of the five O.D.E.'s to the P.D.E. : The differential equations for p and q must continue to give consistent results with the surface calculated - this does happen if the solution is exact, but cannot be hoped for with the noisy data obtained from the image. What one would like to do is continuously monitor whether the current p and q match with the slopes obtained by first differences from points on the current and neighboring characteristics. This is not possible if the characteristics are solved separately and are spreading apart

as well. A later section (4.2.2.1) will explain a method used to continously adjust p and q to remain consistent (derived from the method explained earlier for finding p and q on the initial curve).

At a very minimum, to avoid embarassment one would like to detect when two characteristics approach one another and stop one before they cross. This is easy if the solutions are carried along in parallel, but involves lengthy comparison tests otherwise.

4.2 A PROGRAM SOLVING THE CHARACTERISTICS IN PARALLEL:

Once it had been demonstrated that the equations were correct and a numerical solution possible it was decided to write a second program which would sample the surface of the object more evenly by interpolating new characteristics when needed. Less attention was paid to accuracy in the solution while attempting to be less sensitive to various noise-effects. At the same time an effort to find a more convenient coordinate system produced the much shorter notation and resultant equations described in chapter 2. The solution is achieved by taking all characteristics one step forward at the same time.

4.2.1 THE BASIC DATA STRUCTURE:

The values stored for each point (x, y, z, intensity, p, q and pointers to the previous point on the same characteristic) are here arranged not by characteristic but by 'ring'. A ring is a curve of constant arc-distance from the singular point - i.e. the n^{th} points on all the characteristics form one ring (arranged in counter-clockwise order of the corresponding image points). The complete data-structure is made up of a number of rings, the first of which is the initial curve. As before, individual characteristics may stop for a variety of reasons (s.a. crossing an angular edge) and this causes breaks to appear in the current ring. The break is indicated by a point having a negative intensity, the value being a code for the cause. Some rings thus represent closed curves (e.g. the initial curve) and others more distant from the singular point are broken into sections, the final ring having no active point on it (i.e. positive intensity). Scavenger routines are usually invoked at each solution step and amongst other tasks, compress series of dead points (i.e. negative intensity) into one, since only one is needed to mark a break in the ring.

As we have seen one of the main inducements for using the parallel solution method is to allow interpolation of new

characteristics - this is one of the reasons why the number of points in a ring may change from one to the next and why each point has to have a pointer into the previous ring, indicating which element is its predecessor in the same characteristic. This pointer is -1 if no previous point exists (e.g. on the initial curve or the first point in an interpolated characteristic).

We have seen how characteristics may be terminated causing a break in the ring; it is also possible for a characteristic to disappear, without causing a break, when two characteristics approach too closely. In addition a break can reclose if the points on either side of the break are within the step-size (and pass the crossing-test explained later). With all of this in mind it becomes clear that the data-structure can at times look pretty confused and this has to be remembered when defining a function which interrogates the neighbors of a point (s.a. some sort of difference appproximation).

It was decided to use as data only the coordinates and the slope at each point, because this was sufficient for the uses to be made of the data and also was easily available. For some uses more complicated surface descriptors may be in place, such as the rational function approximations for each

surface-section described by Coon [10]. Usually the increased complexity imposed by such an approach can be side-stepped by rather using a smaller step-size to obtain a finer grid.

It should be noted that the user of constant size steps along the characteristics may produce difficulties on complex objects. For even with smooth surfaces the curves of constant arc-distance from the singular point (the rings) may have cusps. This invalidates the use of difference methods on points along these curves (s.a. are used in subsection 4.2.2.1 and 4.2.2.3). No difficulty was experienced with images of the objects we experimented with. An alternative, which would circumvent this problem, would be the use of steps traversing a constant increment in intensity. This would turn the rings into contours of constant intensity.

4.2.2 EXTRA PROCESSING POSSIBLE:

4.2.2.1 SHARPENING - UPDATING P AND q:

We have already described how one can obtain p(t) and q(t) on the initial curve by solving the set of non-linear equations (subsection 2.3.7):

$$p(t)\, x_t(t) + q(t)\, y_t(t) - z_t(t) = 0$$
$$A(\underset{\sim}{r})\, \phi(I,E,G) - b(\underset{\sim}{r}') = 0$$

In the proof that the solution of the five ordinary differential equations is also a solution of the original partial differential equation, it was stated that the two equations for p and q do in fact continue to give the derivatives of z w.r.t. x and y. When solving a difference equation approximation from noisy data we can expect the solution for p and q to become progressively more inaccurate. Yet the above pair of equations must hold on any path along the surface of the object. In particular one can use them on the curve defined by one ring to determine values of p and q.

For the initial curve we had the additional difficulty that the two equations might have more than one solution and we selected one on the basis of some external knowledge (e.g. that the object is convex near the singular point). We have already assumed that the object is smooth and therefore we will have fairly good values for p and q and cannot get into this difficulty at non-singular points. Even a simple Newton-Raphsen method will suffice to get us more accurate values of p and q.

Let $g(p,q) = p\,x_t + q\,y_t - z_t$

$h(p,q) = \phi(I,E,G) - b/A$

And suppose: $g(p+\delta p,\ q+\delta q) = h(p+\delta p,\ q+\delta q) = 0$

Then ignoring other than first-order terms we have:

$$\begin{pmatrix} g_p & g_q \\ h_p & h_q \end{pmatrix}\begin{pmatrix} \delta p \\ \delta q \end{pmatrix} = \begin{pmatrix} g(p,q) \\ h(p,q) \end{pmatrix}$$

That is:

$$\begin{pmatrix} x_t & y_t \\ p & q \end{pmatrix}\begin{pmatrix} \delta p \\ \delta q \end{pmatrix} = \begin{pmatrix} g(p,q) \\ h(p,q) \end{pmatrix}$$

Here x_t and y_t have to be estimated from difference approximations. One may not want to apply the full correction $(\delta p, \delta q)$. More than one iteration will not be required since p and q are very close to the correct values.

4.2.2.2 INTERPOLATION AND CROSSING TESTS:

When the separation between two neighboring points in a ring becomes greater than 1.5 times the step size along the characteristic, a new characteristic is interpolated. Its x, y, z, p and q values are set to the average of its neighbors while the backward pointer is set to -1. A more

complicated interpolation method can also be used which constructs the line of intersection of the tangent planes at the two neighboring points. It then finds the point on this line closest to the two neighbors and finally uses a point half-way between the point determined previously by the simpler method and this new point (This, for small angles between the tangent planes, is accurate for a spherical surface). This method does not however add significantly to the accuracy of the solution.

If two neighboring points in a section of a ring come closer than 0.7 times the step-size, one is deleted (It is important that this factor be less than 0.75, that is, one half of the factor used in the interpolation decision, or succesive rings on a flat region will have points interpolated on one step, only to be removed on the next, with consequent loss of accuracy).

Finally one wants to stop neighboring characteristics from crossing over each other. Consider the two points a and b on one ring and their successors c and d on the next. The test consists of checking whether c is to the left of the directed line through bd and whether d is to the right of the directed line through ac (Both tests are needed). If either fails, the corresponding characteristic is terminated, causing a

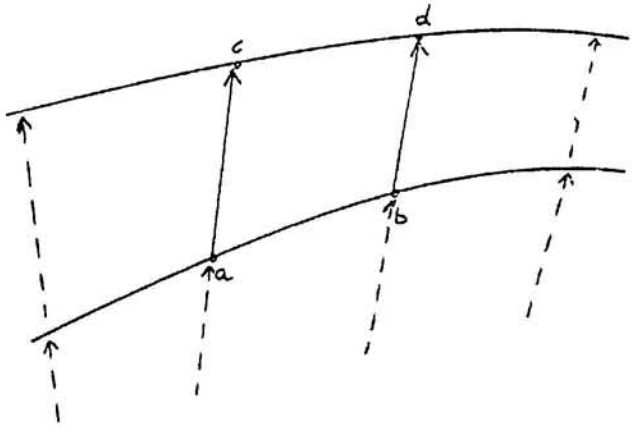


Figure 38: The four points used in the crossing test.

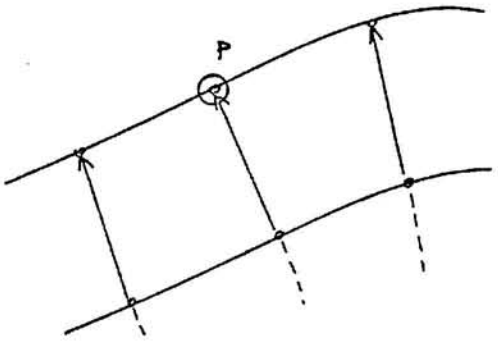


Figure 39: The five neighbors used in determining the intensity gradient at P.

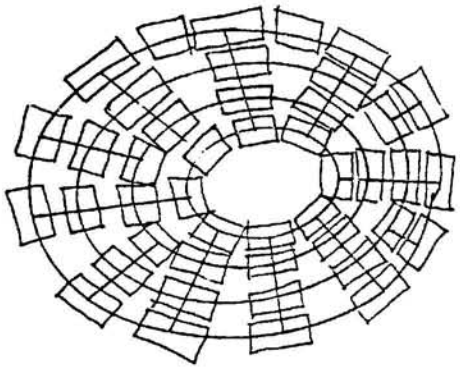


Figure 40: Covering the image with the rasters of points read for each solution point.

break to appear in the ring at that point. The test is equivalent to checking whether the line segment cd falls in front of the line segment ab (and does not cross it). This test is applied across short breaks in rings as well, to stop neighboring section of the ring from crossing over each other.

Care has to be taken if the sections of a ring left all fall on one side of the singular point, since the break then actually encompasses an arc of more than π and crossing tests applied across it will invariable terminate more characteristics on either side of it. This can be avoided if the crossing test is not applied to points whose images fall too far apart (in terms of the projection of the current step-size).

4.2.2.3 OBTAINING GOOD INTENSITY GRADIENTS:

To be more noise-immune than the previous program, a better way had to be found to obtain intensity gradients. Rather than use the intensities at a small raster of points to estimate the local gradient, it was decided to use a difference approximation from intensities measured at neighboring points. Using as many as possible of the

intensities of the point itself and its five immediate neighbors, we can apply a simple least-squares method to estimate the gradient. Some of the points may not exist as explained previously and the characteristic is terminated if less than three points are available or only three which are nearly colinear. Suppose the coordinates of the points are (x'_K, y'_K) (image coordinate system) and the intensities are b_K. We wish to find b_o, $b_{x'}$ and $b_{y'}$, to minimize the following expression:

$$\sum_K (b_{x'} x'_K + b_{y'} y'_K + b_o - b_K)$$

This happens when:

$$\begin{pmatrix} \sum x'^2_K & \sum x'_K y'_K & \sum x'_K \\ \sum x'_K y'_K & \sum y'^2_K & \sum y'_K \\ \sum x'_K & \sum y'_K & \sum 1 \end{pmatrix} \begin{pmatrix} b_{x'} \\ b_{y'} \\ b_o \end{pmatrix} = \begin{pmatrix} \sum b_K x'_K \\ \sum b_K y'_K \\ \sum b_K \end{pmatrix}$$

From $b_{x'}$ and $b_{y'}$ we can find b_x, b_y and b_z by using the camera projection equations of an earlier section (2.7).

For good noise-immunity and some ability to detect surface detail indicating that the solution is invalid, the intensity for each solution point is not read from only one image point. Small tilted rectangular rasters of points are

established around each point of the solution. The one axis of the rectangle is parallel to the base characteristic at that point, and the size is adjusted to correspond to the projection on the image of a square on the object of side-length equal to the step-size. The intensity recorded for a solution point is the average of the intensities read for the points in this raster and the r.m.s./average is used to make the edge-crossing decision. The rasters of all the points in the data-structure almost, but not quite, touch and taken together almost cover the total area of the image explored. This insures that the data is not much affected by pin-holes in the photo-cathode of the image-dissector and that edge crossing can easily be detected, without reducing the resolution.

Both this program and the one discussed in section 4.1 spend more than half their time accessing the image-dissector. Between 20 and 100 intensities are read for each point in the solution, and each access takes about .2 to 1.0 milli-seconds. A complete solution requires from 1 to 5 minutes of real time.

4.2.3 A DOZEN REASONS TO TERMINATE A CHARACTERISTIC:

This is a good place to summarize the reasons for terminating the characteristics. The values printed near the end of a characteristic (derived from the negative intensity code discussed earlier) can be used to index this table.

1. The characteristic has moved out of the field of view of the image-dissector.

2. The r.m.s./average for the intensities read in the raster has become too great, indicating overlap of two objects or an angular joint on one object or some surface detail that is being missed.

3. The intensity has become too low, indicating a shadow region.

4. λ is too large, indicating approach to either another singular point or an ambiguity edge.

5. There are too few neighbors to construct a good estimate of the local intensity gardient.

6. α is too small. α is the Jacobian of the image transformation from z_x and z_y to z_x' and z_y'. This transform becomes singular when $\alpha = 0$. In most cases E will become too small before this happens.

7. A new point was interpolated but both its neighbors were terminated before it could get anywhere.

8. I too small - indicating approach to a shadow edge.

9. E too small - indicating approach to an edge of the object.

10. This characteristic crossed over a neighboring one.

11. It was discovered that this point has a backward pointer to a nonactive point. This is really an error condition and shouldn't normally happen.

12. The intensity is equal to or greater than that measured at the singular point, indicating another singular point or ambiguity edge.

Note that several of these conditions are redundant to ensure that even with an inexact solution at least one will fail at

the right place.

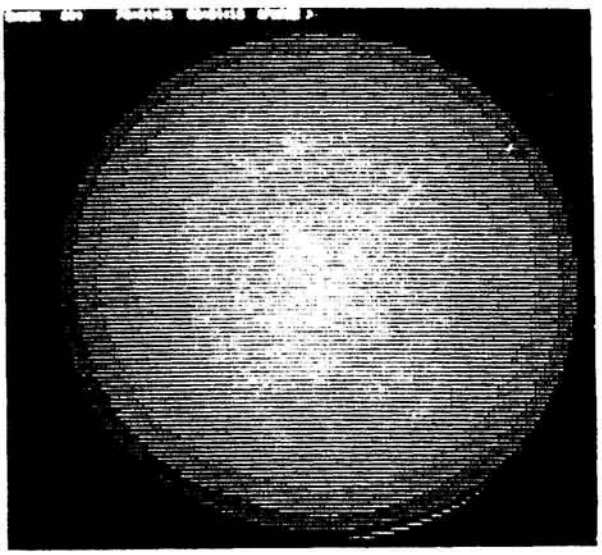
4.2.4 OPERATION OF THE PROGRAM:

4.2.4.1 THE INTEGRATION PROCESS

First the program needs to be given such parameters as the position of the light-source, the distance to the object, focal length of the lens and the step-size to be used in the integration. It then proceeds to find a point of maximum intensity (for some reflectivity functions one needs to search for a minimum). This search can be directed to allow a choice of one of several possible maxima. The program then assumes that this point of maximum intensity is a singular point and that the object is convex at this point (in some cases we would like to assume it to be concave). After constructing an initial curve (a small circle) around the singular point, it proceeds to read the intensities at the corresponding image points. The non-linear equations for p and q on this curve are then solved iteratively.

All intensities are normalized w.r.t. the intensity at the singular point unless the surface has a specular component. In the latter case, the intensities on the initial curve are

FIGURE 41A



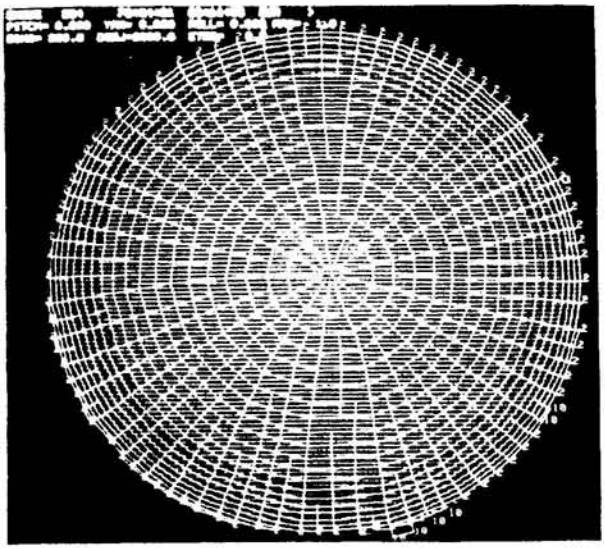
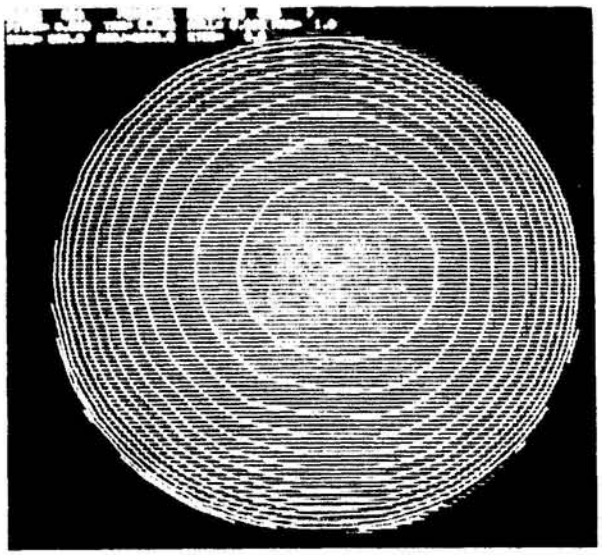
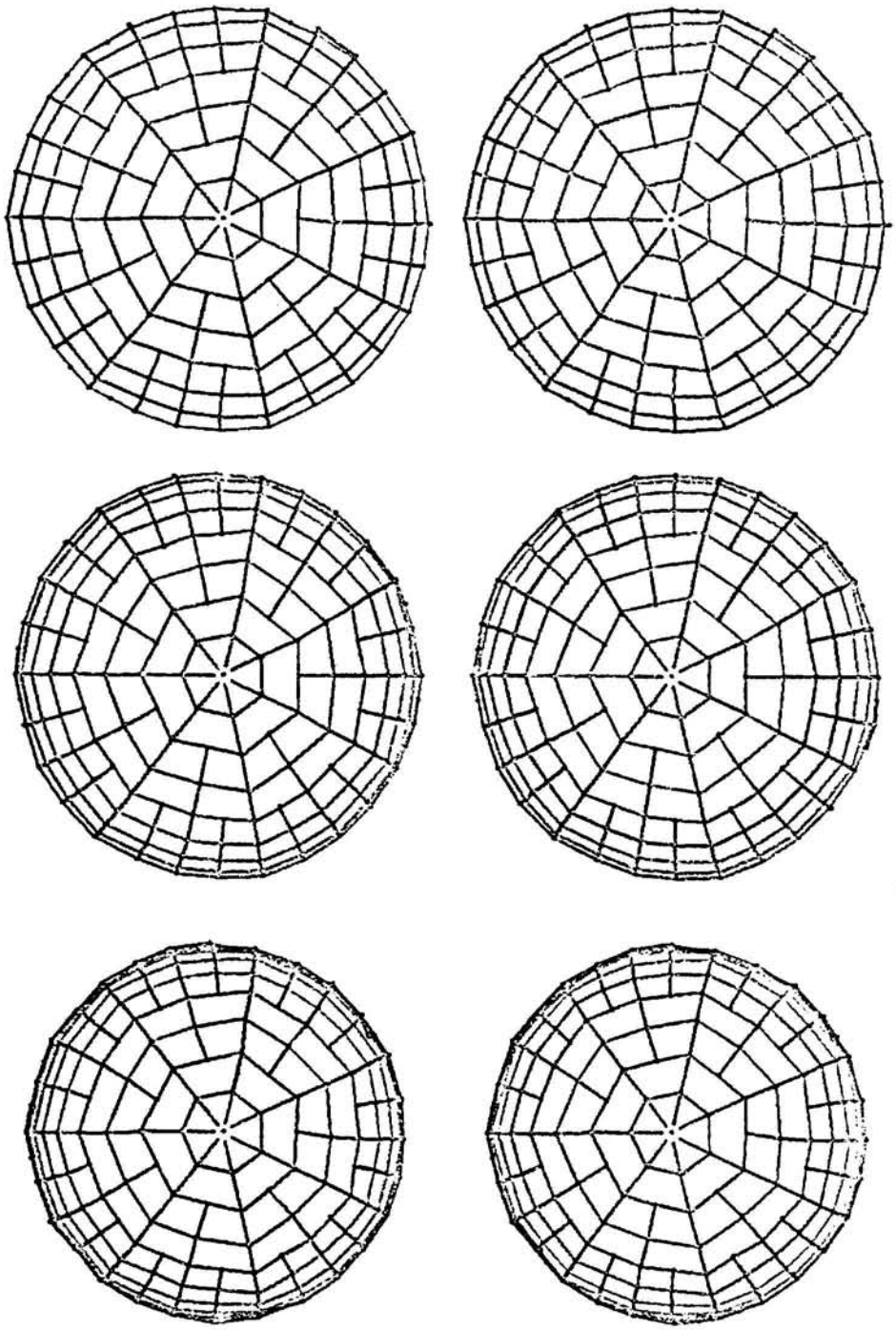


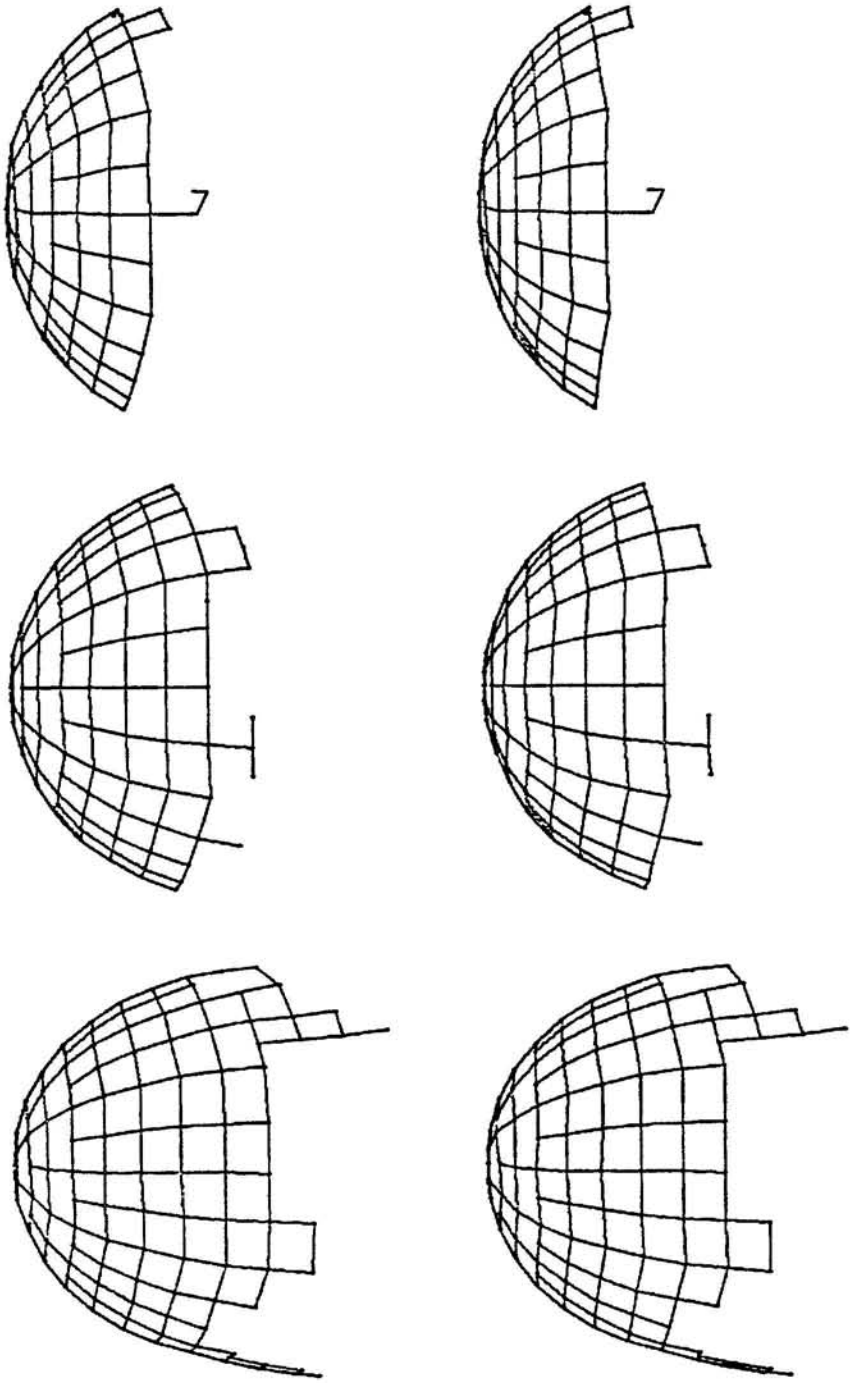
FIGURE 41B

FIGURE 41C



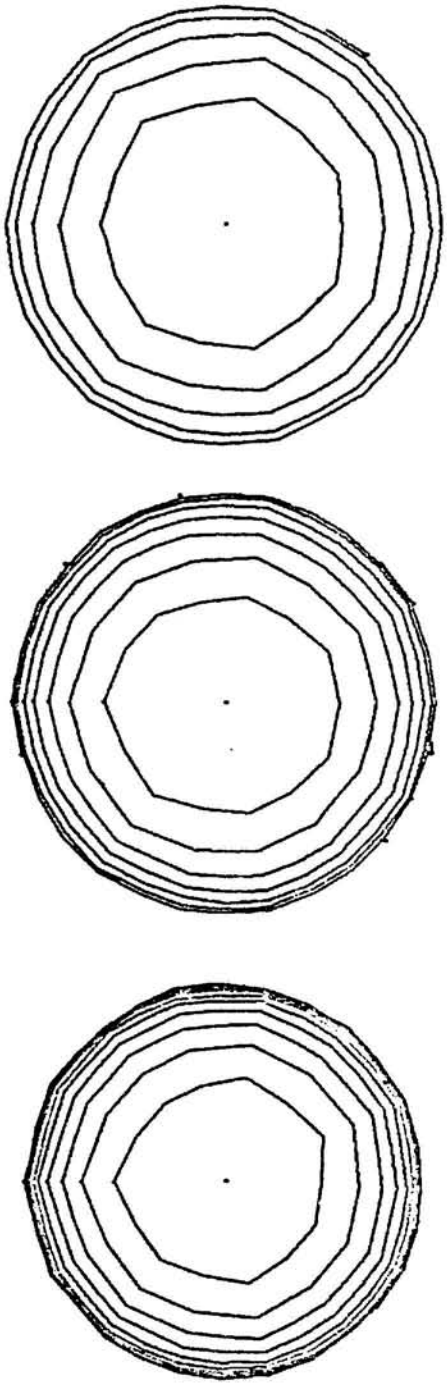


Figures 42 A, B, C: Stereo-pairs of solutions produced by new SHADE for disc-shaped, spherical and bullet-shaped objects (actually spheres with make-up applied).



Figures 43 A, B, C: Stereo-pairs of same solutions as in previous figures, rotated 90^{o}.

SHADE 314 70:05:06 06:25:27 BULLET 1



Figures 44 A, B, C: Contour maps of same solution as in previous figures.

used to establish a normalization value (The specular reflectivity is too variable for use in normalization). It is assumed that the initial curve has been chosen large enough fo fall outside the region of strong specular reflection.

For each step in the parameter s, the following procedure is then carried out:

1. For each point calculate the normal ($\underset{\sim}{n}$), the incident vector ($\underset{\sim}{r_i}$) and the emittance vector ($\underset{\sim}{r_e}$). From these obtain the derivatives $I_{\underset{\sim}{n}}$, $E_{\underset{\sim}{n}}$ and $G_{\underset{\sim}{n}}$.

2. Calculate ϕ_I, ϕ_E, ϕ_G and hence $\phi_{\underset{\sim}{n}}$.

3. Then obtain F_p, F_q and λ.

4. Add $(\delta x, \delta y, \delta z)$ to (x,y,z) to get the point on the next ring for each characteristic. Here $(\delta x, \delta y, \delta z) = \lambda(F_p, F_q, pF_p+qF_q)$.

5. Interpolate new points where the points in the new ring are too far apart and delete points where they are too close together. Produce breaks where characteristics have crossed over adjacent characteristics.

6. Now read the intensities for all the points. Terminate those characteristics with points of very low intensity or high r.m.s/average.

7. Calculate $b_{x'}$, $b_{y'}$ for all those points for which enough neighbors exist. From these values obtain b_x, b_y and b_z by the projection equations.

8. Now use $\underset{\sim}{n}$, $\underset{\sim}{r}_i$ and $\underset{\sim}{r}_e$ to calculate $I_{\underset{\sim}{r}}$, $E_{\underset{\sim}{r}}$ and $G_{\underset{\sim}{r}}$.

9. Next use ϕ_I, ϕ_E and ϕ_G to calculate $\phi_{\underset{\sim}{r}}$.

10. Then obtain F_x, F_y and F_z.

11. Add $(\delta p, \delta q)$ to (p,q) to obtain p and q for the uninterpolated points on the new ring. Here $(\delta p, \delta q) = \lambda((-F_x - pF_z), (-F_y - qF_z))$.

12. Interpolate p and q for the new points.

13. Sharpen up the values for p and q on all points in the new ring.

14. Garbage-collect various items, such as series of points with negative intensity.

15. Stop if no points with positive intensity remain.

It should be apparent where the various tests for terminating the characteristics fit into the above schema. The simple Euler method for solving the differential equations could be replaced by a Runge-Kutta method with increases in running time of a factor of two, but little improvement in accuracy. The sharpening method, on the other hand, is very cheap and contributes substantially to accuracy.

4.2.4.2 OTHER PROCESSING AVAILABLE

As explained before, the data-structure is displayed as it is generated and can also be viewed from different angles when completed. In addition a mode exists where the mapping from three-space to the display surface is not performed by the projection explained earlier, but a simple map from a rectangular area on the image-dissector to a rectangular area on the display surface. This is particularly valuable for overlaying the solution on an intensity modulated display of what appears in the image. This aids greatly in debugging since it is easy to pinpoint such problems as starting the solution from an inappropriate maximum in intensity.

A number of other displays can be produced to aid in setting up the image-dissector. Prodigous amounts of detailed print-out can be generated during a solution process and a more parsimonious listing of the final data is available. It is possible to substitute synthetic data (with selectable amounts of noise) for the image-dissector input as a repeatable way of checking out the program and to tide over those days when the image-dissector is being repaired! The data can be written to and read from the disk and tape.

The stereoscopic display has to be viewed with an appropriate pair of lenses which are not always handy. For this reason a routine was provided which produces a contour map from the data. This map is produced by first listing the intersections of all the lines in the data structure (from point to point in each characteristic, as well as from point to point in each ring) with the selected contour planes. The intersections are then sorted on contour plane. Within each contour plane the following process is applied repeatedly until no points are left:

> Pick a point and find the closest neighbor within a 'reasonable' distance. 'Reasonable' distance is defined to be 1.5 times the step-size used in the solution. Now another point is selected closest to the new point also

within a reasonable distance and so on until no more can be found. The point chosen at each step may not be the first in the chain so constructed (which would close the loop) unless no other points are available. Also the line-segments connecting sucessive points may not make angles of more than $\pi/2$ with one another. The points are removed from the data as they are used in generating the contour except the very first point (to allow for the eventuality of closing the contour).

The distances are usually weighted with the dot-product of the new segment with the previous segment, to give preference to continuation of contours in the direction of the last segment used. The method generates good contours where the data is complete and smooth, and does fairly well otherwise.

4.2.5 INSENSITIVITY TO IMPERFECTIONS IN THE SENSOR:

This program is not quite as accurate as the one that solves the characteristics serially (mostly because of the simple method for solving the differential equations numerically), but vastly superior in its behavior when faced with noisy data. Most of the improvement is due to the better way of obtaining intensity gradients and to the use of the lateral

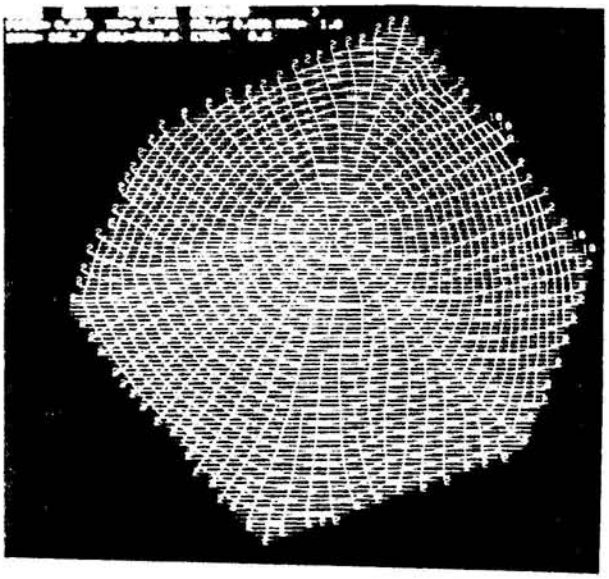
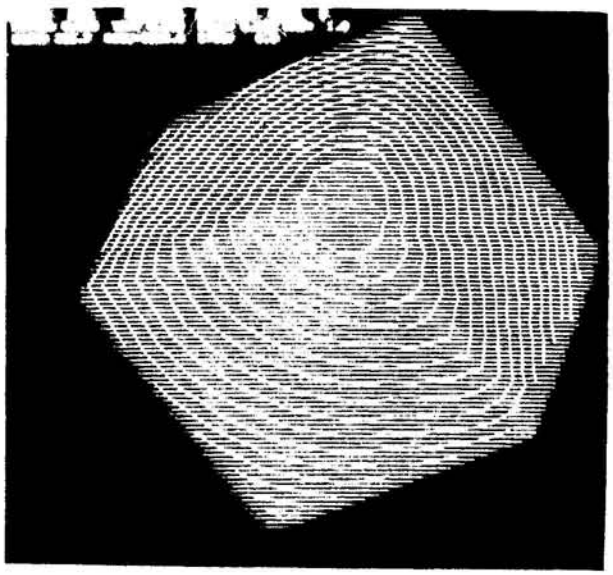
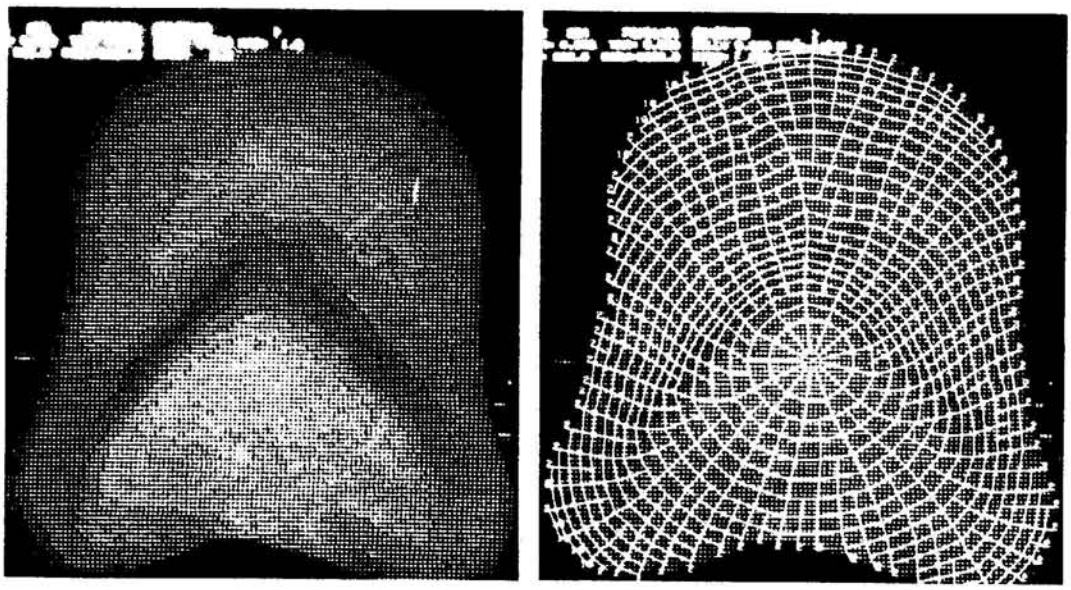


FIGURE 45A

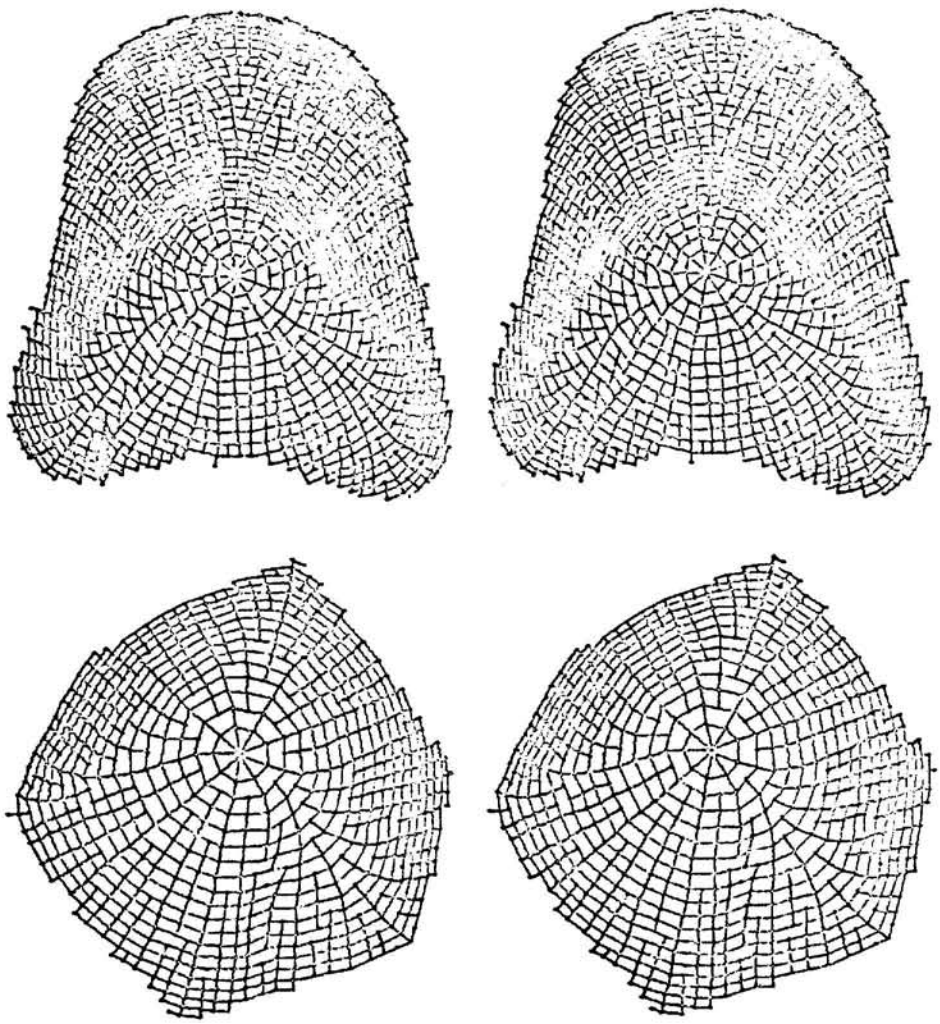
FIGURE 45B





FIGURES 46 A,B,C

SHADE 314 70:05:07 10:26:31 CUBE 2



Figures 47 and 48: Stereo-pairs of solutions obtained for the plaster object and the cube with rounded corners.

```
SHADE    314      70:05:07   13:25:52  FUM      2
PITCH= 0.000   YAW=-1.300   ROLL= 0.000 MAG=   3.0
DIMG= 114.0   DOBJ=1000.0   EYES=   68.0
```

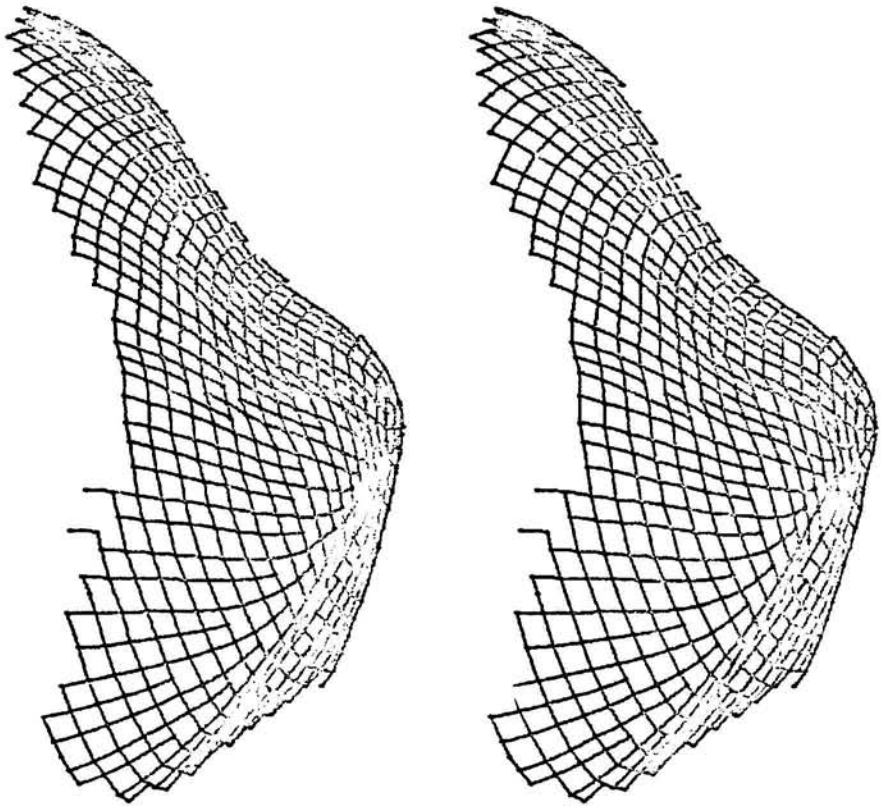


Figure 49: Stereo-pair of side-view of solution obtained for the plaster object.

SHADE 314 70:05:07 11:44:55 CUBE 2

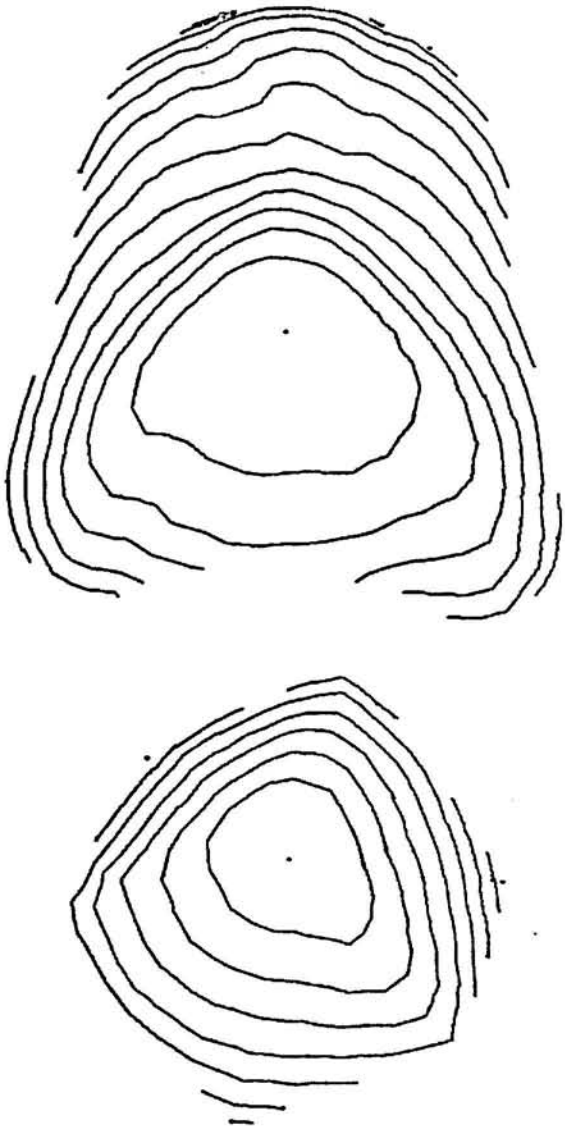


Figure 50 and 51: Contour maps of the solutions obtained for the plaster object and the cube with rounded corners.

connection between the characteristics. The difference approximation for the intensity gradients uses a support area about six times as large as the one used by the least squares approximation of the first program.

Distortions in the imaging device 'merely' produce distortions in x and y, while non-uniformities in the sensitivity will affect p and q and hence z. The only effect of low resolution will be that some edges will not be noticed and the solution erroneously continued across them.

4.3 A NOSE-RECOGNITION PROGRAM:

To illustrate one use of the shape-from-shading method, it was applied to a simple recognition task. Although there is great interest in face-recognition [12] (partly because there is a practical use for it), it was decided to tackle a sub-problem - that of nose-recognition. In principle, face-recognition could be carried out by repeating the process explained here for not only the nose, but the chin, forehead and the two cheeks. Transparencies of noses, rather than real noses were used because they are always ready and do not move during the minute or so it takes to determine the shape. To avoid having to determine the reflectivity function for

skin as a function of all three angles, special lighting conditions were employed. The light-source was placed near the camera and the reflectivity function as a function of the incident angle determined from the transparencies taken. This meant that no separate determination of the non-linearities in the photographic process was needed.

4.3.1 MODIFICATIONS TO THE BASIC PROGRAM REQUIRED:

A few minor changes and additions had to be installed in the main program for this task. Most prominent amongst these is the procedure used to normalize the intensities read from the image. Because of the strong specular component of highly variable nature, the singular point could not be used for this normalization. The specular component in the transparencies not only varies from person to person and time to time but depends on the exposure used, since it usually is bright enough to saturate the film. Normalization was thus carried out w.r.t. an intensity derived from that measured on the initial curve, which was assumed to be outside the specular region.

```
SHADE   298      70:04:29  06:33:18  GOOD   2
PITCH= 0.000  YAW= 0.000  ROLL= 0.000 MAG=  0.7
DIMG= 140.0  DOBJ=1000.0  EYES=   0.0
```

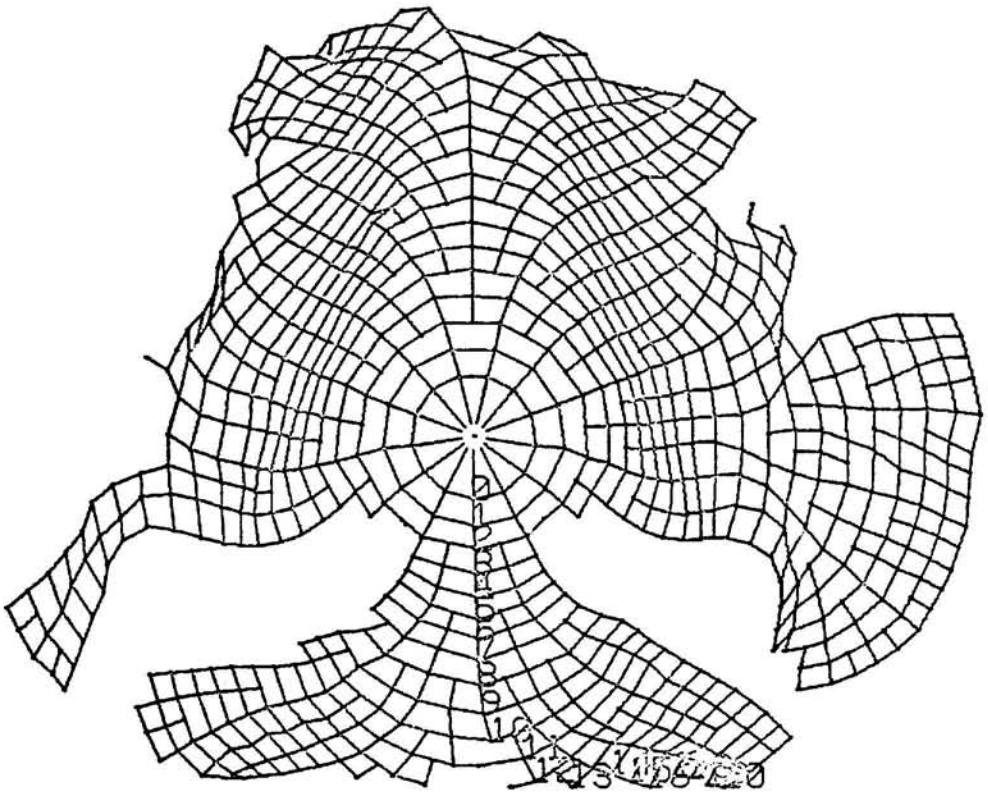


Figure 52: Solution obtained for a nose.
Note gaps left by the breaks caused by
the nostrils.

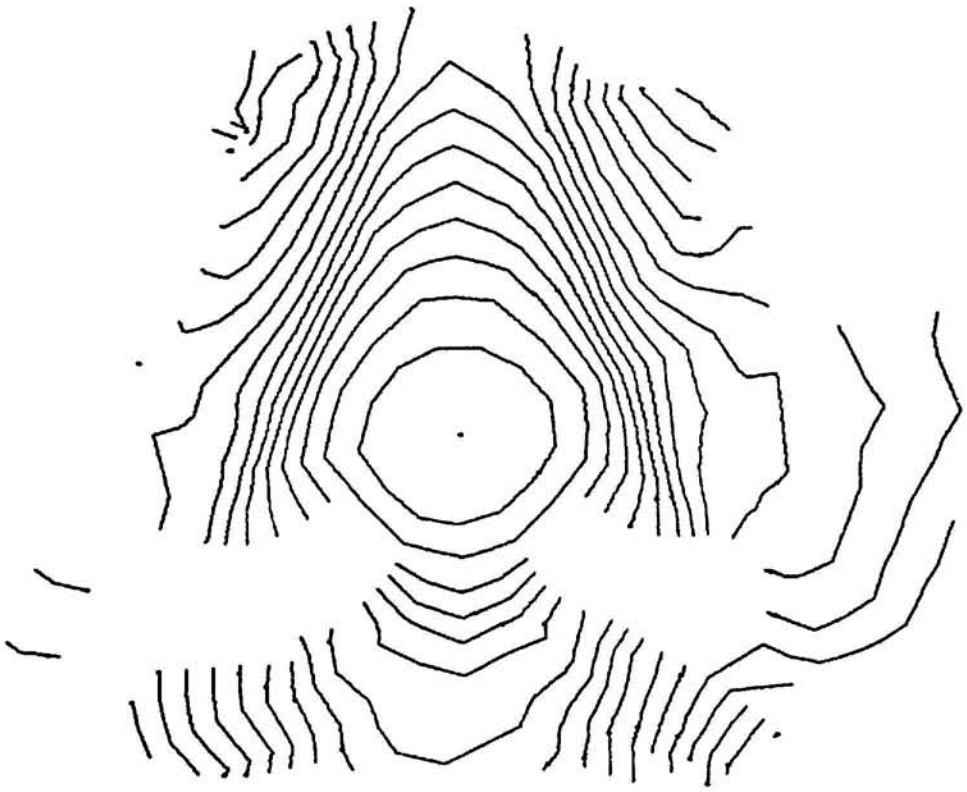


Figure 53: Contour map of solution obtained for a nose.

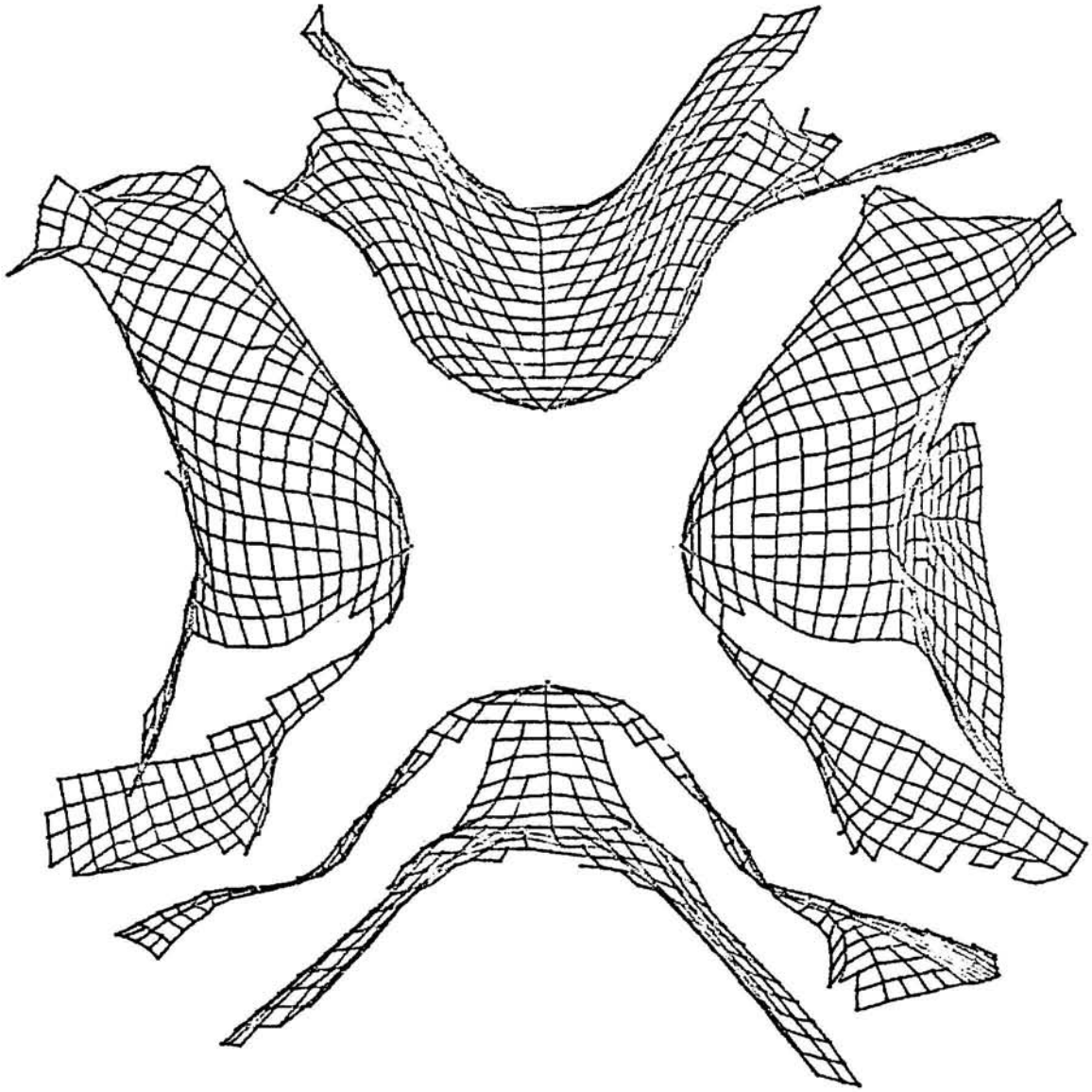


Figure 54: Four views of solution obtained for a nose.
(With some hidden lines eliminated).

```
SHADE   314      70:05:06  07:18:00  GOOD   3
PITCH= 0.000  YAW= 0.000  ROLL= 0.000 MAG=  2.7
DIMG= 114.0  DOBJ=1000.0  EYES=  68.0
```

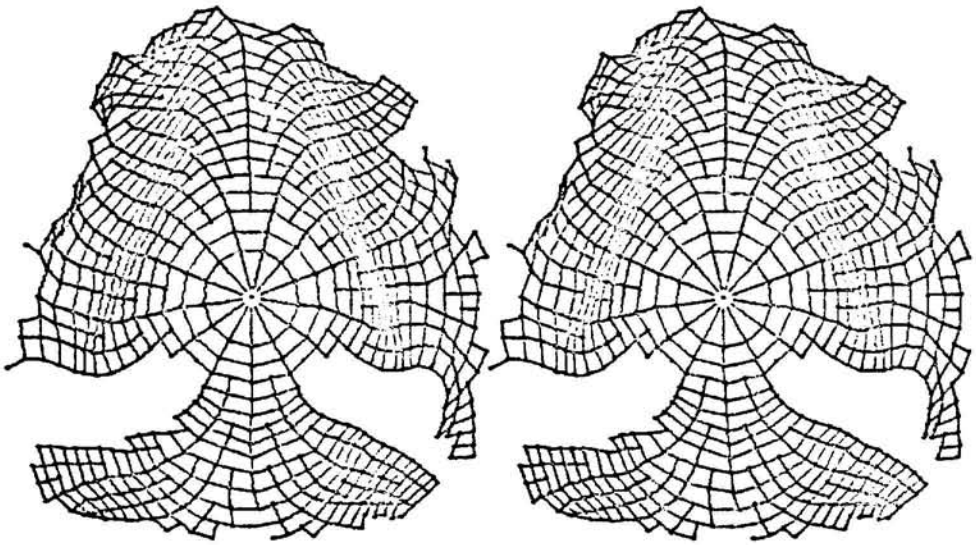


Figure 55: Stereo-pair of solution obtained for a nose.

4.3.2 NORMALIZATION PROCEDURE:

In order to simplify photographing the subjects, it is necessary to make some decisions about which factors one is going hold fixed and which are to be taken care of by some normalization in the program. Altough it is possible to hold the head in a standard position by means of a bite-bar, it is inconvenient and it is preferable to let the program take care of small head-rotations. The distance from the camera to the subject on the other hand is very easy to determine and therefore no normalization of size was used. For pictures of the whole head such size normalization would be fairly accurate, whereas it cannot be for images of the nose alone which does not present sharp features to take measurements of.

The rotational normalization procedure to be described can handle quite large ($<\pi/6$) rotations in both pitch (rotation about an ear to ear axis) and roll (rotation about a tip-of-nose to back-of-head axis). Yaw (rotation about a top-of-head to throat axis) is restricted by the requirement that almost all of the surface of the nose should be visible. For some noses this restricts the rotation to fairly small angles - of course this presents no problem when taking the photograph.

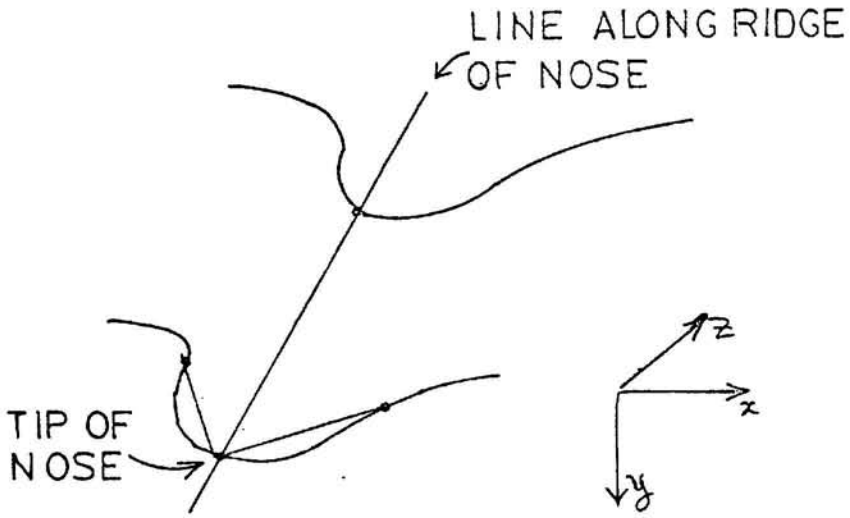


Figure 56: Illustration of rotational normalization procedure.

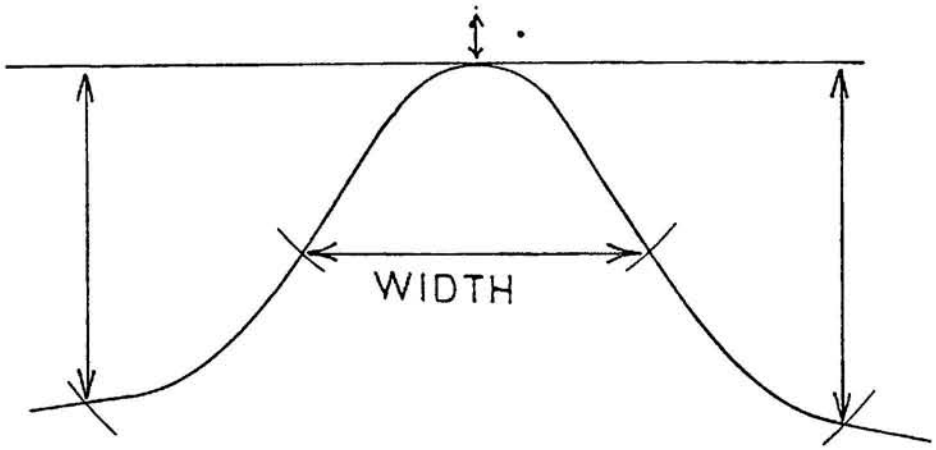


Figure 57: Illustration of parameters abstracted from one horizontal contour through the nose.

Independence of rotation is achieved by means of a routine which establishes the orientation of the shape calculated and then rotates it into a standard position. In addition the parameters in the final comparison procedure where chosen to be independent of small remaining errors in the orientations. The orientation of the shape calculated is estimated from two horizontal contours through the nose, one passing through the tip of the nose, the other higher up on the ridge. These contours of course are only defined as sequences of points where the characteristics and rings pass through each plane.

The most forward points defined by these contours are calculated by fitting a parabola to the three points with lowest z coordinates. For each of the two contours we get one such forward point, connecting them we obtain a line which runs approximately along the ridge of the nose. This line is rotated into a standard position (Lying in the y-z plane and leaning $\pi/6$ from the vertical).

The lower contour (through the tip of the nose) is also used to estimate rotation about the vertical axis. The two points on this contour at a given distance from the most forward point define two angles w.r.t. the z-axis. The desired rotation is one half of the difference of these two angles.

The three angles so determined are small and can thus be treated independently. The rotation of the shape is performed about the center of the spherical cap used to determine the initial curve, i.e. a point just inside the tip of the nose. The whole process is repeated iteratively three times. The errors remaining are almost always less than *0.01* radian (*0.5* ^{o}). It was found that using only the few points indicated to determine the rotation was quite satisfactory, although better accuracy is no doubt obtainable if the calculation employed averages over several points.

4.3.3 COMPARISON PROCEDURE:

After the data has been brought into a standard orientation, we would like to abstract a small number of parameters which contain most of the information for comparison purposes. A rather arbitrary decision was taken to use estimates of the distance of the ridge of the nose from the standard line (In the y-z plane and leaning $\pi/6$ from the vertical), the width of the nose about half-way down to the cheek and the depth of the cheek from the ridge of the nose. These quantities where measured for each of five horizontal contour planes, the lowest through the tip of the nose, the highest a bit below the saddle point (the bridge between the eyes). The fifteen

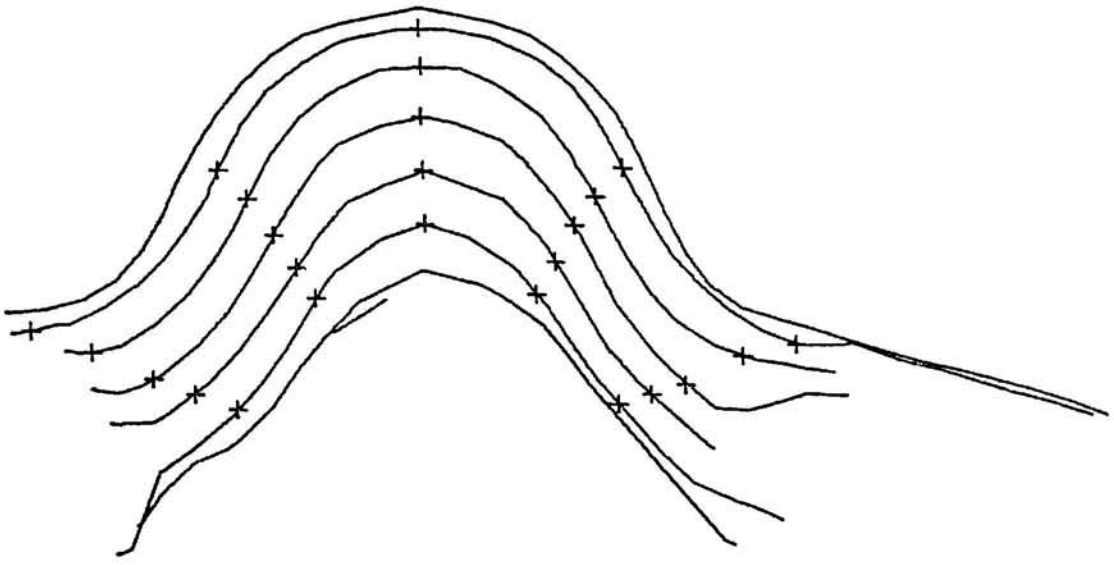


Figure 58: The points on the 5 contours used to abstract the fifteen values describing this nose.

values so obtained are the only data used in the final comparison procedure.

The distance down the side of the nose from where measurements of the width of the nose are taken varies with the contours, going from some large value for the plane passing through the tip of the nose to one-half that value for the highest contour. The distance at which the depth of the cheek is measured is twice that at which the width of the nose is measured and thus also varies from contour to contour. The depth of the cheek is the average of the depth obtained on the left side and that obtained on the right. The fifteen measurements obtained for each transparency are stored in a table together with the number of the transparency.

The purpose of the comparison procedure is to establish if any of the stored measurements match those obtained from a new transparency. To determine this, a pseudo-distance is calculated (in the 15-dimensional vector space), between each stored vector and the new vector. The pseudo-distance is a weighted r.m.s. of differences in coordinates [12], where the weights are proportional to the standard deviation observed for that coordinate.

$$d^2 = \sum_i (x_i'' - x_i')^2 / \sigma_i^2$$

where d is the pseudo-distance, x_i'' and x_i' the components of the two vectors, and σ_i the standard deviation of the i^{th} component. The uncertainty in the depth to the cheek is greater than that in the width of the nose, for example, and it therefore has a lower weight than the latter. This procedure gives a comparison test which is in some sense optimal [12].

No doubt other comparison procedures and other choices of parameters would have been equally useful; in particular it soon become apparent that fewer than 15 parameters would have been equally as selective. The point is that once one has data as complete as a full description of the shape, almost any method will work and it is not even necessary to display great sophistication in one's use of statistics.

4.3.4 RESULTS OF THE NOSE-RECOGNITION PROGRAM:

15 transparencies of 12 noses were used in this experiment. The pairs of transparencies for the three noses which were photographed twice differed in camera to subject distance, head rotation and exposure. A total of 30 shapes were

FIGURES 59 A,B,C

FIGURES 60 A,B,C

calculated, two each for those noses of which only one transparency was available (they differed because of the noisy nature of the data). For each shape so determined, the rotational normalization was applied and the 15-tuple description abstracted. The pseudo-distances between all pairs of 15-tuples were then determined.

The pseudo-distance between 15-tuples averaged to the following (the units are about 0.3 mm's r.m.s.):

1. Between transparencies of different noses - 10. (range 2.4 - 15.3)

2. Between transparencies of the same nose - 2. (range 1.4 to 2.5)

3. Between shapes calculated from the same transparency 1. (range 0.1 to 2.1)

In all cases the distance from a given shape to a related shape was less than a quarter of the distance to a unrelated shape. Simply looking for the smallest pseudo-distance (and checking whether it is fairly small), thus gives an effective recognition procedure for this small data-set. It is clear that for a much larger data-set unique identification would

Figure 61

	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B	7A	7B
1A	0	1	10	9	8	8	11	11	6	6	10	11	10	10
1B	1	0	11	10	8	8	11	11	6	6	10	11	10	11
2A	10	11	0	2	11	11	14	14	7	8	14	15	13	12
2B	9	10	2	0	11	11	13	13	7	7	13	14	13	12
3A	8	8	11	11	0	2	8	8	4	4	6	7	8	9
3B	8	8	11	11	2	0	9	9	4	4	7	8	9	10
4A	11	11	14	13	8	9	0	0	10	8	3	3	3	4
4B	11	11	14	13	8	9	0	0	10	8	3	3	3	4
5A	6	6	7	7	4	4	10	10	0	2	9	10	9	9
5B	6	6	8	7	4	4	8	8	2	0	7	8	8	8
6A	10	10	14	13	6	7	3	3	9	7	0	1	4	5
6B	11	11	15	14	7	8	3	3	10	8	1	0	5	6
7A	10	10	13	13	8	9	3	3	9	8	4	5	0	2
7B	10	11	12	12	9	10	4	4	9	8	5	6	2	0

Table of pseudo-distances between some of the shapes calculated. Pairs 2, 3 and 5 are each of two different transparencies of one nose, while the other pairs are each two shapes calculated from one transparency. The units of distance are about 0.3 mm r.m.s. .

be more unlikely without improving the accuracy in the solution and a detailed analysis of which parameters to abstract for optimal recognition. It would however always be possible to separate out some small subset of the total stored set of nose-descriptions with very high probability that the nose looked for will be in this set. Bledsoe [12] uses the ratio of the size of this subset to the size of the complete stored set as a measure of the effectiveness of the recognition procedure.

Repeating the operations we described here for the other large frontal planes (planes with normal parallel to the z-axis), one would obtain a face-recognition procedure. It is very likely that the subsets of all stored face-decriptions determined by applying the above method to cheeks, chin, forehead and nose in turn will have only a small intersection. This is not to say that other information about the face, not obtainable from the shape-from-shading method could not add to the accuracy of such a procedure. It must be pointed out that some of the feature points used in previous attacks on the face-recognition problem are not defined by sharp discontinuities (for example the tip of the nose) and could best be obtained from a description of the shape.

The restriction about the positioning of the light-source could be removed if one took the trouble to measure the reflectivity function in more detail and either recorded the positioning of the light-source or worked out in detail a method for finding the single light-source from the shadows in the image (which should not be very difficult since we know the approximate shape of the object we are looking at). The full face-recognition problem was not tackled since it would require a great deal more work without further illustrating the method of determining shape-from-shading. Also it will be noted that the study involved a small set of noses - a study with a large data-set would contribute little more to the understanding of the method.

Some of the difficulties encountered when determining the shape of noses are perhaps worth mentioning. Firstly, most noses are not completely visible from any given point of view. Most notably the underside (between the nostrils) is frequently not visible, and often a small area on the side of the nostrils is also hidden. This forced a choice of parameters which did not depend on these areas. Naturally the information of whether these areas are visible could in itself be useful in the recognition procedure if it could be reliably determined. In fact our program does not, because of the combination of poor resolution in the image-dissector

and the simple-minded edge detector. This could be circumvented by placing the light-source slightly above the camera, thus ensuring that there always is a narrow shadow below the nostrils.

When the solution is erroneously continued across an edge (such as that above the nostrils), a second undesireable effect appears because of the sharpening procedure. The incorrect coordinates of the points calculated after the edge is crossed have some effect on their neighbors due to this and thus decrease the accuracy of the solution obtained nearby.

Another problem is that some noses have not one, but two closely spaced tips (probably because of the underlying cartilage consisting of two symmetrical parts). This causes the characteristic growing from one of these peaks towards the other to stop, since it is approaching another singular point. A simple solution consists of chosing the radius of the initial curve large enough to completely include both singular points. Finnally one finds that some noses (particularly those belonging to females) have very low ridges near the eyes, making it difficult to determine a meaningful value for the width of the nose at that point.

It should be noted that the reflectivity function was not determined with great precision and no account was taken of its variation from person to person. It was not important that the shape calculated was very close to that of the nose from which the image was taken, but rather that differences in the shapes of noses should show up as differences in the calculated shapes and that shapes determined from transparencies of the same nose should be similar. If the images were all produced with the heads in the same rotational position, the distortions would have made no difference at all. For the small head rotations encountered, the effect of the relatively minor distortions was very small.

4.4 SUMMARY AND CONCLUSIONS:

After defining the reflectivity function, an equation was found relating the intensity measured in the image of a smooth opaque object to the shape of the object. This equation was then shown to be a first-order non-linear partial differential equation in two unknowns and the equivalent set of five ordinary differential equations was derived. A number of especially simple cases were discussed, in particular applications to lunar topography and

the scanning electron microscope. Methods were described for obtaining the auxiliary information required (e.g. the reflectivity function) and how to avoid the need for an initial known curve on the object. Of importance too is the method demonstrated for continuously updating p and q (sharpening) as the solution progresses.

The half-dozen or so other depth-cues were ignored here to allow a comprehensive treatment of shading. The analytical approach to the problem of determining shape from shading was developed to demonstrate that an exact solution is possible and to determine just what the limitations of this approach are. This is not to say that a more heuristic, approximate approach does not have its merits too for certain types of objects [14]. It was decided to produce a program to allow experimentation with the solution method because many ideas in the field of artificial intelligence and visual perception are of little value until they can be tried on real data. Fortunately an image-dissector was available to provide input of image intensities to the computer.

Two programs were presented, one solving the O.D.E.'s for the characteristics sequentially, the other in parallel. Advantages of the latter approach were found to be several. Finally this latter program was adapted to provide input for

a nose-recognition procedure.

It has been made apparent that shading is valuable as a monocular depth-cue altough it may not be as accurate as some others. It must be emphasized that no claim is made that people employ this depth-cue in the same way. It may be that the human visual system does not actually determine the shape in three-space and if it does so it is likely that it uses a different method. However there will be many similarities between the two systems (e.g. in the errors they make) because they utilize the same data.

4.4.1 SUGGESTIONS FOR FUTURE WORK:

1. It would be instructive (but very time consuming) to measure many reflectivity functions and see how many fall into the pattern of a matt component, approximately varying with cos(i), plus a specular component. If it could be shown that most real reflectivity functions fall into this class, the method presented would be more useful since it could determine approximate shapes without knowing much more about the reflectivity function.

2. It may be possible to find more simplifying conditions s.a. the ones found with certain lighting conditions, positions of the light-source and special reflectivity functions.

3. Other solution methods may be found, or modifications to the integration method might increase the accuracy. Perhaps a difference method on a fixed grid could be found which somehow gets around such problems as that of ambiguity edges.

4. One could study the two related problems of finding the reflectivity function, given the shape of the object and the light-source distribution and finding the light-source distribution given the reflectivity function and the shape.

5. Further study of certain types of inconsistencies and their use is indicated. Here for example we find the problem of deciding whether certain faces in an image of several polyhedra could consistently belong to one object.

6. Some effort could be directed towards implementing more fully some of the ideas developed theoretically here,

s.a. shadow bridging, handling multiple sources and multiple singular points.

7. Expanding the nose-recognition program into a full face-recognition program would increase its usefulness.

8. One could study in more detail how people use the depth-cue of shading and how bad animals are at it. Perhaps one can get a better clue as to whether people develop a three-dimensional model of the object from the shading or if they use the shading information in some other way.

9. There are probably a few more loose ends such as the problem of how to start the solution if no convex or concave singular points are available. Can one do anything at all with saddle points (even though they can camouflage themselves to be indistinguishable from simple convex or concave singular points)?

10. In addition to interpolation, is it reasonble to extrapolate? That is, can one generate new characteristics next to a solution sheet to explore new areas. In particular when a break appears in a solution surface can it be patched-up later?

11. More methods will have to be found to deal with the three-dimensional structure once one has determined it.

12. As pointed out earlier, the use of constant size steps along the characteristics may not be ideal (remember that we can adjust the step-size by choosing a different λ). One particularly attractive idea would be to use steps corresponding to constant intensity change in the image. This would turn the rings into constant intensity contours, rather than curves of constant arc-distance from the singular point.

13. Many objects have surfaces whose reflectivity cannot be described by a function of three angles, or are so specular that our methods are of little avail. One might try to discover methods of dealing with such objects. Examples are chrome car-bumpers, translucent wax, hair and a glass of water.

5. REFERENCES:

1. V. P. Fesenkov: 'PHOTOMETRIC INVESTIGATIONS OF THE LUNAR SURFACE' 1929 Astronomochheskii Zhurnal 5 (Translated by Redstone Scientific Information Centre).

2. J. van Diggelen: 'A PHOTOMETRIC INVESTIGATION OF THE SLOPES AND HEIGHTS OF THE RANGES OF HILLS IN THE MARIA OF THE MOON' July 1951 - Bulletin of the Astronomical Institute of the Netherlands.

3. D. E. Willingham: 'THE LUNAR REFLECTIVITY MODEL FOR RANGER BLOCK III ANALYSIS' November 1964 - Technical Report 32-664 Jet Propulsion Laboratory.

4. T. Rindfleisch: 'PHOTOMETRIC METHOD FOR LUNAR TOPOGRAPHY' March 1966 - Photogrammetric Engineering.

5. D. R. Garabedian: 'PARTIAL DIFFERENTIAL EQUATIONS' 1964 - John Wiley.

6. R. D. Richtmeyer and K. W. Morton: 'DIFFERENCE METHODS FOR INITIAL VALUE PROBLEMS' 1957 - John Wiley

7. R. W. Hamming: 'NUMERICAL METHODS FOR SCIENTISTS AND ENGINEERS' 1962 - McGraw Hill.

8. K. C. Kelly: 'A COMPUTER GRAPHICS PROGRAM FOR THE GENERATION OF HALF-TONE IMAGES WITH SHADOWS'. November 1969 - University of Illinois.

9. B. K. P. Horn: 'THE IMAGE DISSECTOR 'EYES'' August 1969 - A.I. Memo 178 - M.I.T.

10. S. A. Coons: 'SURFACES FOR COMPUTER-AIDED DESIGN OF SPACE FIGURES' February 1968 - Mechanical Eng M.I.T.

11. P. R. Thornton: 'THE SCANNING ELECTRON MICROSCOPE' November 1965 - Science Journal

12. W. W. Bledsoe: 'MAN-MACHINE FACIAL RECOGNITION' August 1966 - Panoramic Research Report 22 - Palo Alto

13. C. Engelman: 'MATHLAB' August 1968 - IFIP Congress [page B91 to B96] - Edinborough

14. L. J. Krakauer: 'COMPUTER ANALYSIS OF VISUAL PROPERTIES OF CURVED OBJECTS' June 1970 - Electrical Eng M.I.T.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA – R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Massachusetts Institute of Technology
Project MAC

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
None

3. REPORT TITLE

Shape from Shading; A Method for Obtaining the Shape of a Smooth Opaque Object From One View

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Ph.D. Thesis, Department of Electrical Engineering, June 1970

5. AUTHOR(S) *(Last name, first name, initial)*

Horn, Berthold K. P.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
November 1970	198	14

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Nonr-4102(02) b. PROJECT NO.	MAC TR-79 (THESIS)
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
None	Advanced Research Projects Agency 3D-200 Pentagon Washington, D.C. 20301

13. ABSTRACT

A method will be described for finding the shape of a smooth opaque object from a monocular image, given a knowledge of the surface photometry, the position of the light-source and certain auxiliary information to resolve ambiguities. This method is complementary to the use of stereoscopy which relies on matching up sharp detail and will fail on smooth objects. Until now the image processing of single views has been restricted to objects which can meaningfully be considered two-dimensional or bounded by plane surfaces.

... A number of applications of this method will be discussed including one to lunar topography and one to the scanning electron microscope. In both of these cases great simplifications occur in the equations. A note on polyhedra follows and a quantitative theory of facial make-up is touched upon.

An implementation of some of these ideas on the PDP-6 computer with its attached image-dissector camera at the Artificial Intelligence Laboratory will be described, and also a nose-recognition program.

14. KEY WORDS

Artificial Intelligence
Depth Cues
Visual Perception
Machine Vision
Image Processing

DD FORM 1 NOV 65 1473 (M.I.T.)

UNCLASSIFIED
Security Classification