AD-A008 786

ILLUMINATION FOR COMPUTER-GENERATED IMAGES

Bui Tuong Phong

Utah University

Prepared for:

Rome Air Development Center
Advanced Research Projects Agency

July 1973

ILLUMINATION

FOR

COMPUTER-GENERATED IMAGES

by

Bui Tuong Phong

COLOR ILLUSTRATIONS REPRODUCED IN BLACK AND WHITE

July 1973

UTEC-CSc-73-129

This research was supported in part by the University of Utah Computer Science Division and the Advanced Research Projects Agency of the Department of Defense, monitored by the Rome Air Development Center, Griffiss Air Force Base, New York 13440, under contract F30602-70-C-0300.

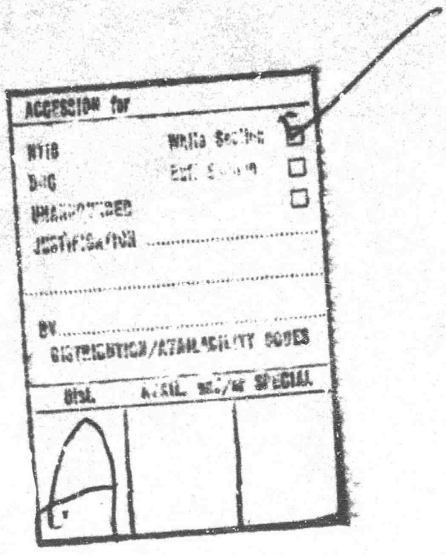


TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

ABSTRACT*

This report describes a new model for the shading of computer-generated images of objects in general and of polygonally descibed free-form curved surfaces in particular. The shading function is determined by a linear interpolation of the curvature of the surface. It takes into consideration the physical properties of the materials of which the surfaces are made. By applying the fundamental laws of optics, highlights due to the specular reflection of the light are simulated, and other existing shading problems are overcome. This results in computer-generated images of increased realism.

A large number of sample pictures are provided to give a pictorial comparison of the new shading process with past methods. Finally, pictures of simple real solids, such as cylinders, spheres and cones, are compared with those generated by the computer.

* This report reproduces a dissertation of the same title submitted to the Department of Electrical Engineering, University of Utah, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

CHAPTER I

INTRODUCTION

This research describes a new approach to the production of shaded pictures of solid objects. In the past decade, we have witnessed the development of an increased number of systems for the rendering of solid objects by computer. The main computational problem has been the elimination of hidden parts of the objects. Higher quality computer-generated images are desirable, and although several attempts to improve this quality have been made, most of the effort has been spent in the search for fast hidden-surface removal algorithms. With the development of these algorithms, the programs which produce pictures are becoming remarkably fast, and we may now turn to the search for economical algorithms for enhancing the quality of these images. Also hardware facilities now permit real-time simulation, which opens the door to a wide area of applications, such as artificial intelligence, architecture, arts, and other areas where realism is highly desirable.

I. 1 Historical account of present methods.

The progress made during the last decade in the technique of displaying three-dimensional objects on a cathode ray tube has been dramatic. Since 1963 when Roberts [1] at the MIT Lincoln Laboratory first performed removal of hidden

lines, a large spectrum of new and original algorithms for displaying visible surfaces have been developed. Recently Sutherland et al [2] summarized and classified these algorithms in "A Characterization of Ten Hidden-Surface Algorithms." While each of these methods is different in its approach to the problem of hidden-surface removal, most of the authors were not primarily concerned with the visual quality of the generated images. Beautifully shaded pictures have been produced but at prohibitive computation costs for real-time dynamic picture.

The image quality depends directly on the effectiveness of the shading algorithm, which in turn depends entirely on the method of modeling the object. Two principal methods to describe the objects are commonly used:

1. Surface definition using mathematical equations.
2. Surface approximation by planar polygonal mosaic.

Several systems are implemented which remove hidden parts for mathematically defined curved surfaces (MAGI [3], Comba [4], Weiss [5], Mahl [6]). With these systems, exact information at each point of the surface can be obtained, and the resulting computer generated pictures are most realistic. However, the class of possible surfaces is restricted, and the computation time needed to remove the hidden parts and to do the shading is very high. Significant developments in applications to restricted surfaces have been accomplished. At the University of Utah, Mahl has worked on this problem for quadratic patches. Currently Catmull [7] is implementing a hidden-surface algorithm for cubic surfaces, and Clark [8] is designing a interactive real-time display system for free-form B-spline surfaces.

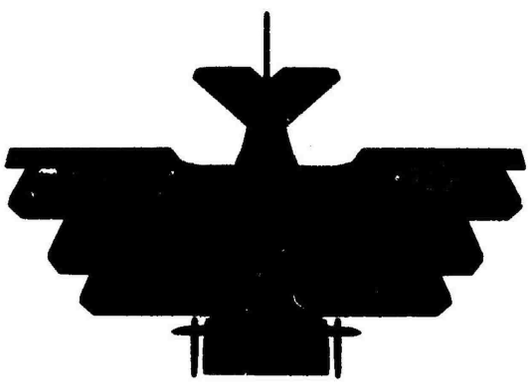
A simple, rough method of representing curved surfaces and objects of any arbitrary shape is to approximate the surfaces with small planar polygons, that is, a piecewise linear approximation. This type of representation has the advantage that it avoids the problem of solving higher order equations that is typical of mathematically curved surface approaches. Most of the hidden-surface algorithms developed in the last few years are based on this type of model structure. Among them, Galimberty [911], Kubert [10], and Loutrel [11] worked on line drawings, and Romney [12], Warnock [13], and Watkins [14], introduced shaded pictures. Starting from the Utah work, Bouknight [15] and Kelley [16] worked on systems for shaded picture display with shadows and movable light sources. Also Appel [17] has developed a system for machine renderings of solids with shadows.

The common goal of these algorithms is a high efficiency which would allow a hardware implementation for real-time display of complex objects. This goal has been reached by the two following systems. An early hardware implementation of a hidden-surface algorithm done by Rougelot et al [18] at General Electric for the N.A.S.A. was based on the polygonal description of curved surfaces. The latest hardware realization for a real-time display of complex objects has been accomplished by Watkins at Evans & Sutherland Computer Corporation.

Most of the earlier implementations did not seem to concentrate on developing the shading algorithm. Warnock was the first to try to investigate in depth the notion of color, hue, brightness, and specular reflection of light, applied to computer-generated images. Warnock implemented these ideas in his system and proposed a model for display of shaded pictures in which the light source is located at the eye, and curved surfaces are approximated with planar

(a)

Without depth simulation

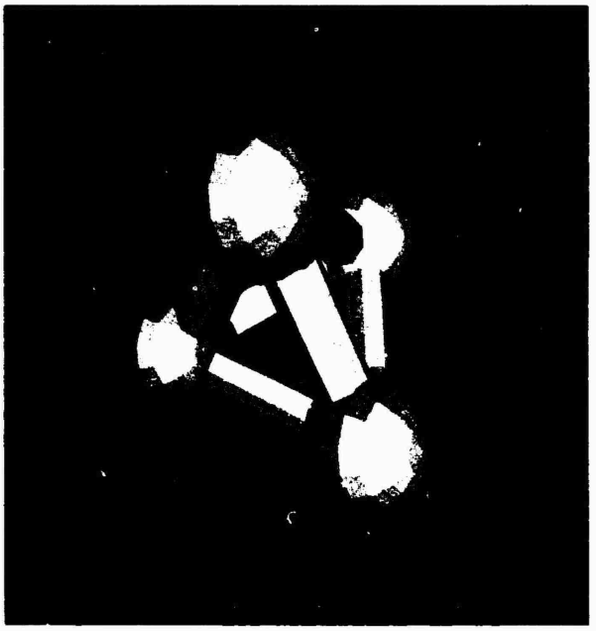


(b)

Figure 1.1 With depth simulation

(a)

Without specular reflection



(b)

Figure 1.2 Warnock model

polygons. With this model, the shade of each displayed polygon is uniform all over its surface. It is the sum of two functions: the hue of the object and the specular reflection of the light. As the three-dimensional object is projected into the screen space, all of its parallel surfaces will have the same shade on the generated pictures. The feeling of depth is then destroyed. In order to restore the illusion of depth , the shade of each point of the object is also made a function of the distance between the point and the light source.

To illustrate Warnock's model, several pictures are generated and presented here. The two pictures on Figure 1.1 represent an airplane, with three parallel wings. On the left picture, the three wings are indistinguishable, because they are painted with the same shade. On the right picture, generated with the addition of the distance parameter, a difference of shade is created between these wings, giving the illusion of depth. The pictures on Figure 1.2 illustrate the specular reflection of light, using Warnock's model. The polygons facing toward the observer are brighter than the ones just adjacent to them, due to the specular reflection of light.

Although the described model gives increased realism to the pictures, several drawbacks of this model still exist. By assuming that the light source is located at the eye, this model fails to apply to most applications in real life. Take the example of an airplane landing on a runway: if the light is located at the eye of the pilot, the oblique angle of illumination leaves the runway completely dark as seen from the cockpit of the pilot when the plane is touching the ground! The second drawback of this system is that the "facetted" approximation disturbs the smoothness of the curved surface.

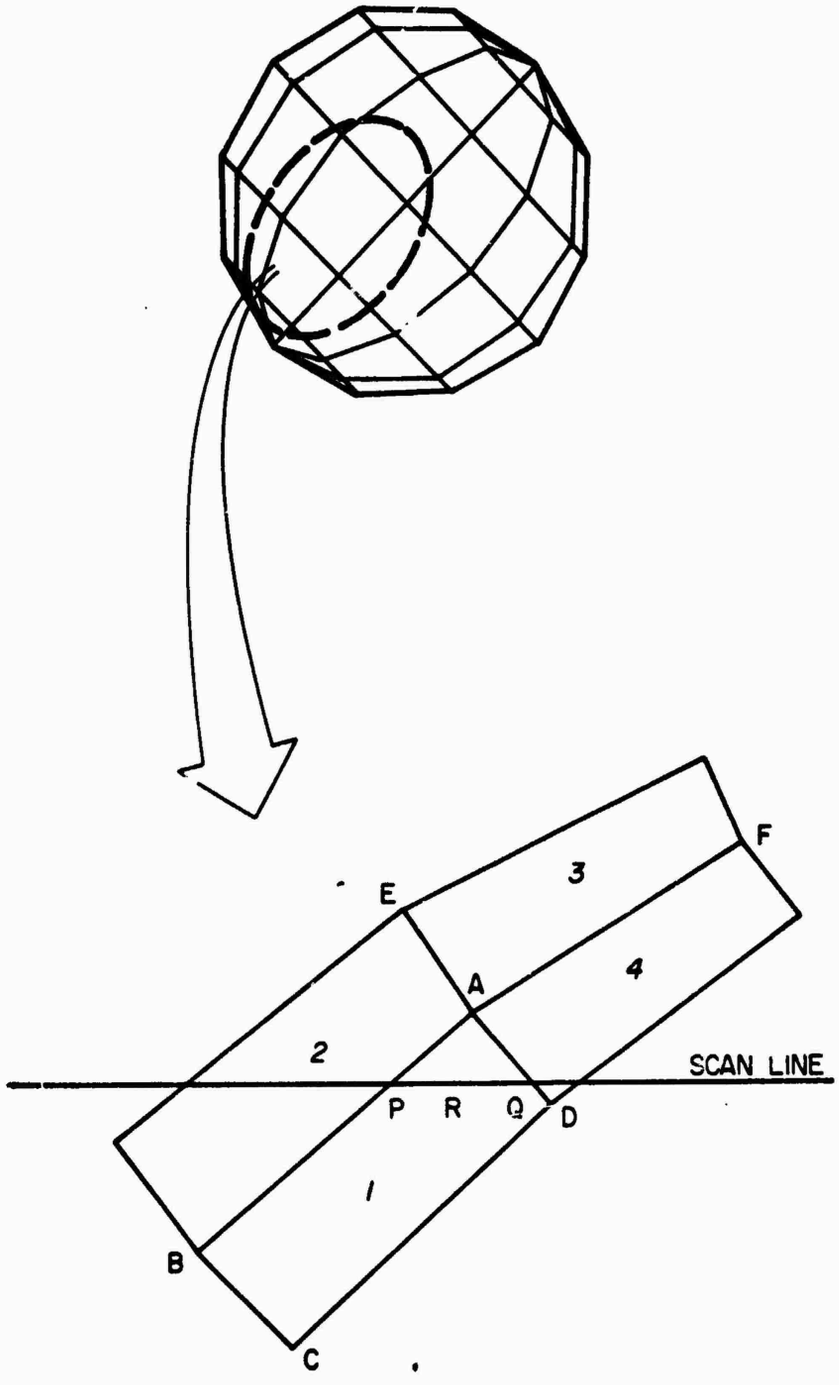
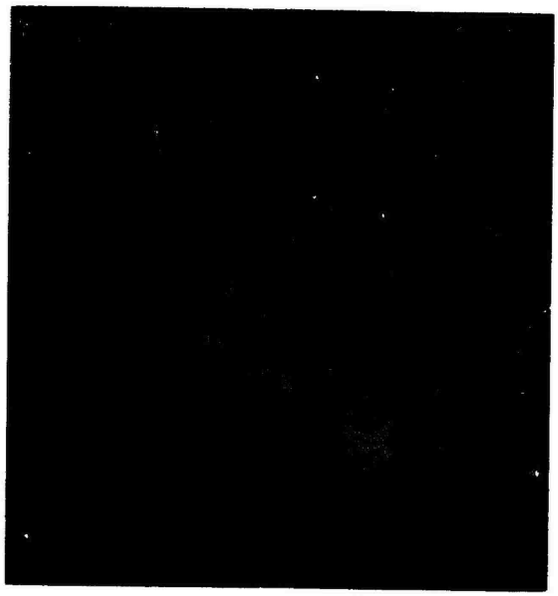


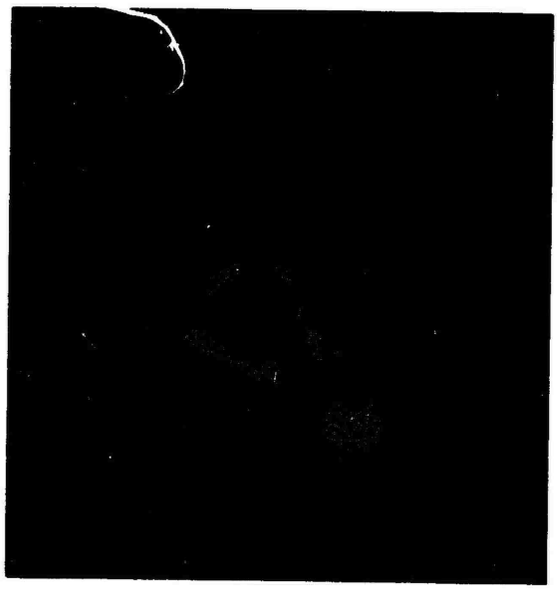
Figure 1.3 Computation of the shading at point R using Gouraud method: two successive linear interpolations:

across polygon edges: P between A and B, Q between A and D
across the scan line: R between P and Q



(a)

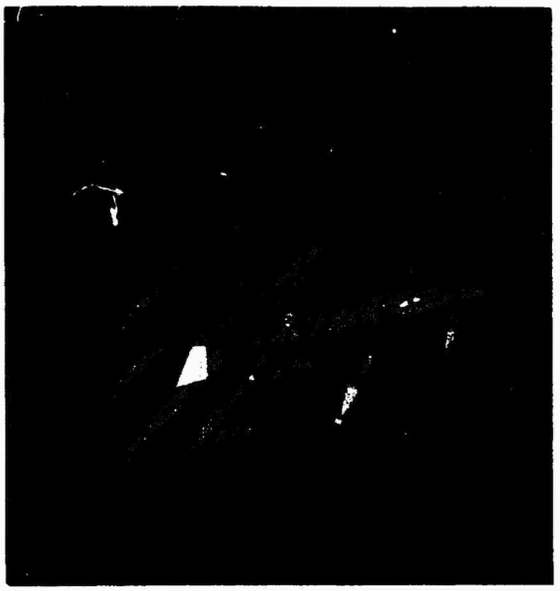
Faceted surface



(b)

Gouraud shading

Figure 1.4



(a)

Faceted surface

(b)

Gouraud shading

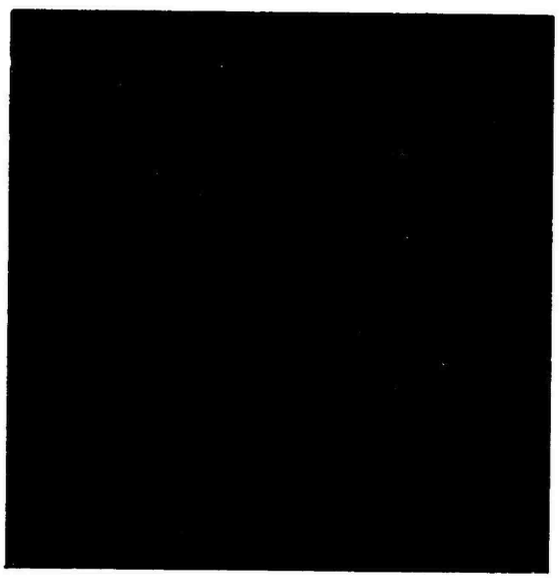
Figure 1.5 B-58 airplane

To solve some of these problems, Gouraud [19] introduced a new way to represent curved surfaces. Instead of having the information about the curvature of the surface for each facet, Gouraud keeps this information at each vertex of the surface. From the curvature, a shade intensity is computed and retained. When the surface is displayed, this shade intensity is linearly interpolated along the edge between adjacent pairs of vertices of the object. The diagram on Figure 1.3 gives an example of the determination of the shade at a point on the surface, using the Gouraud method. This very simple method gives a continous gradation of shade over the entire surface, which in most cases restores the smooth appearance

.

Figures 1.4 and 1.5 show the difference between the Gouraud smooth shading technique and the previous methods. However with Gouraud's method, the the subjective discontinuity of shade at the edges still disturbs the smoothness of the surface, and additional problems arise when simulation of specular reflection of light is attempted. These problems will be discussed later.

Independent of the work done at Utah, Newell, Newell, and Sancha [20] at the Computer-Aided Design Centre in Cambridge, England, presented some ideas on creating highlights. From observations in the real world, they found that highlights are not only created directly by the light source, but also by the reflection of light from other objects in the scene. This is especially true in the case of objects made of highly reflective and transparent materials. In Newell's model, curved surfaces are approximated with planar polygons. Since the light source can be arbitrarily located, the shading function for every polygon depends on the orientation of the polygon with respect to the light source and the line of sight. Unfortunately, the ability to generate highlights using either



(a)

Simulation of transparent objects

(b)

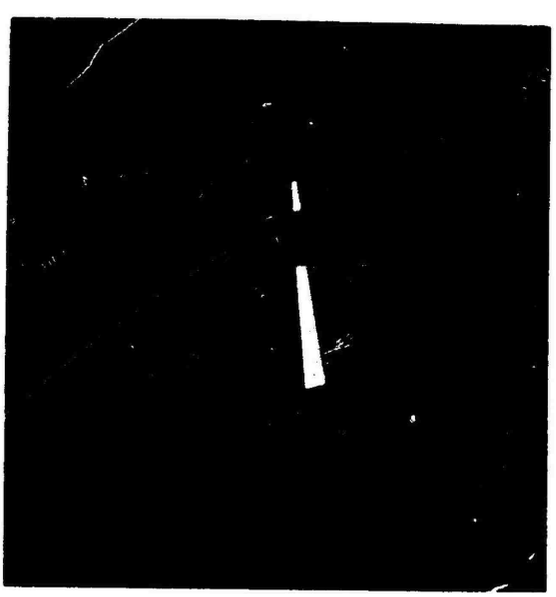


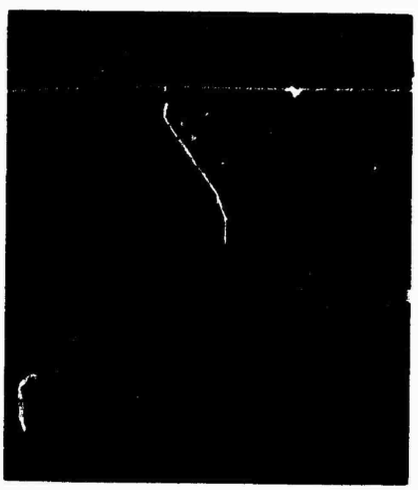
Figure 1.6

Newell's or Warnock's methods is severely limited due to the inability to vary light intensity over the surface of any single polygon. Although the Gouraud technique improves this ability, it still does not render realistic highlights as explained in the following section. Figure 1.6 shows a transparent cup on the chess board and a glass bottle using the Newell model of shading.

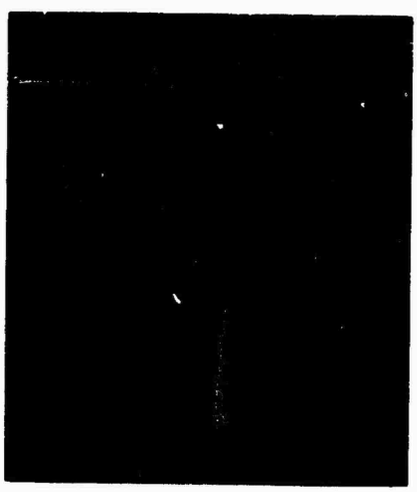
I. 2 Problems with Existing Methods.

With the introduction of the Gouraud smooth shading technique, the quality of computer-generated images allows representation of a large variety of objects with great realism. However, certain problems still exist, one of which is the apparent discontinuity across polygon edges. Highlights on surfaces rendered with a high component of specular reflection are often inappropriately shaped, since they depend upon the disposition and shape of the polygons used to approximate a curved surface and not upon the curvature of the object surface itself. Finally, the shading of a surface in motion in a computer generated film has annoying frame to frame discontinuities due to the changing orientation of the polygons describing the surface. The shading algorithms are not invariant under rotations.

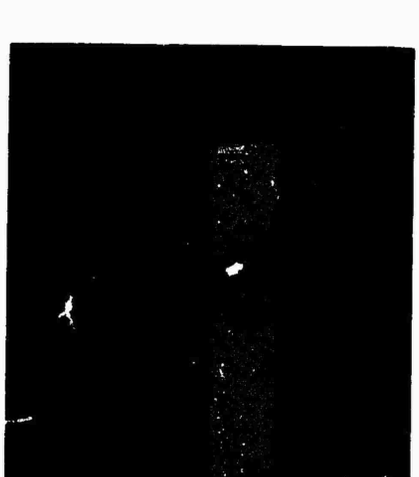
The problem of discontinuities across polygon edges will be studied in detail in the next section about Visual Perception and the Mach Band effect. Figure 1.7 shows a cylinder approximated by three, four, eight, and twelve polygons respectively, to illustrate this phenomenon. As demonstrated by these pictures, the discontinuities are less apparent with a larger number of polygons, but they do not disappear completely.



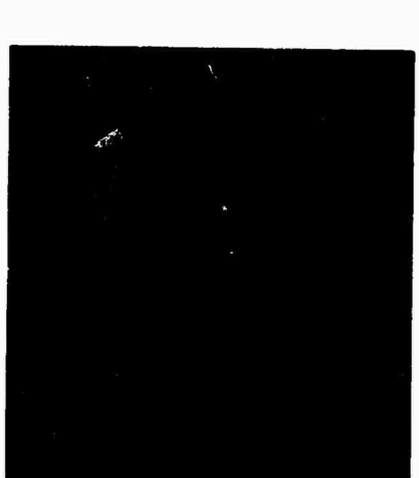
(a) 3-sided model



(b) 4-sided model



(c) 8-sided model



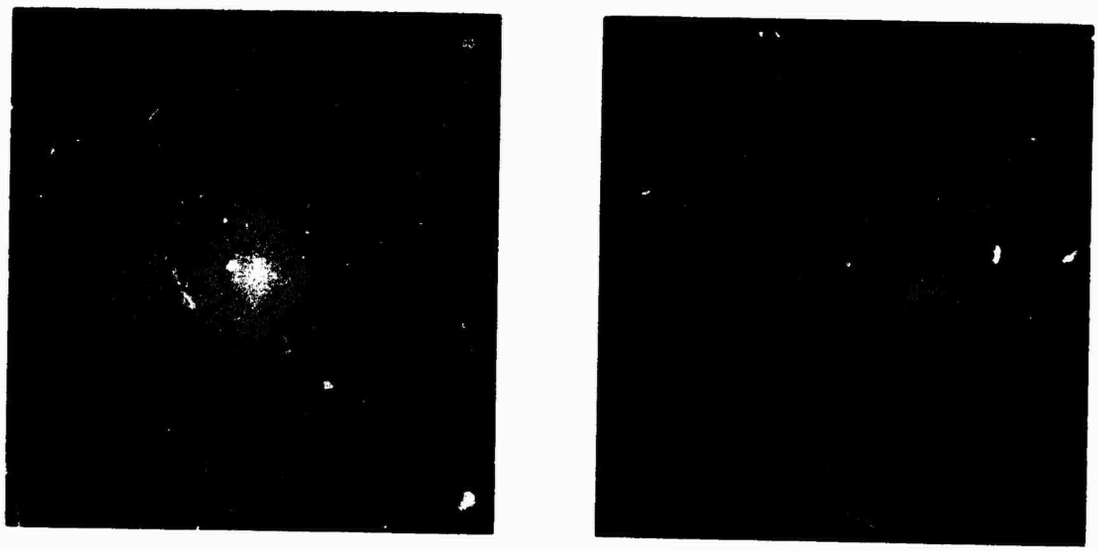
(d) 12-sided model

Figure 1.7 Different cylinder models

The deformation of the shape of the highlights using Gouraud smooth shading technique can be illustrated with the following example (see Figure 1.8a): A sphere is approximated with a certain number of triangular and square facets (Figure 1.9). Assuming that the light source is located at the eye of the observer, Figure 1.10a shows a projection of the sphere and the eye onto a plane. The vectors Ni are the directions of the normals to the sphere at vertices Pi (i=1,...,n). By simple laws of optics, the reflected light received by the observer from any point on the sphere depends on the angle made between the direction of sight and the reflected light vector at that point. For example, at point P1 of the Figure 1.10a, the specular reflected light intensity is a function of the angle (EP2,R2), where E is the common position of the eye and the light source, R2 is the direction of the reflected light at point P2. If this angle is greater than ninety degrees, it means that the direction of the reflected light is away from the eye; there will be no specular reflected light received by the eye from this point.

In Figure 1.10b , the size of the dots is proportional to the intensity of the specular reflected light received by the eye from this point. At point P2 all the reflected light is received by the eye, and therefore the amount of reflected light at this point is the largest. From points P2, P3, P6 and P7, a part only of the reflected light strikes the eye, while there is no other specular component from the rest of the vertices.

When the sphere is displayed using the Gouraud scheme of linear interpolation of the shade function for smooth shading of the curved surface, the following phenomenon occurs: in representing with a contour the points which have the same amount of specular reflected light, the shape of this contour is not a



Light source at observer's point of view Light to the right of observer

Figure 1.8 Highlights simulation with Gouraud shading

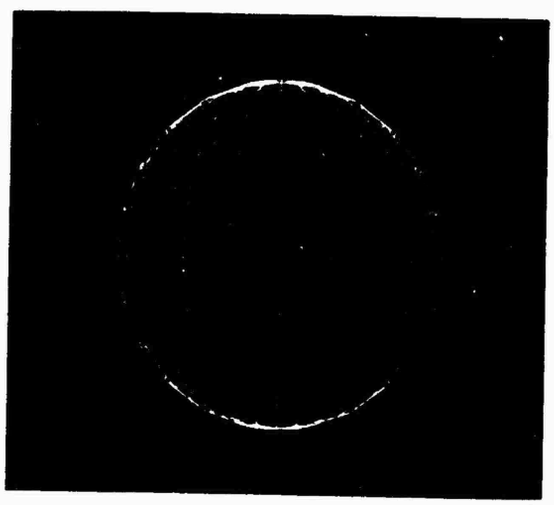


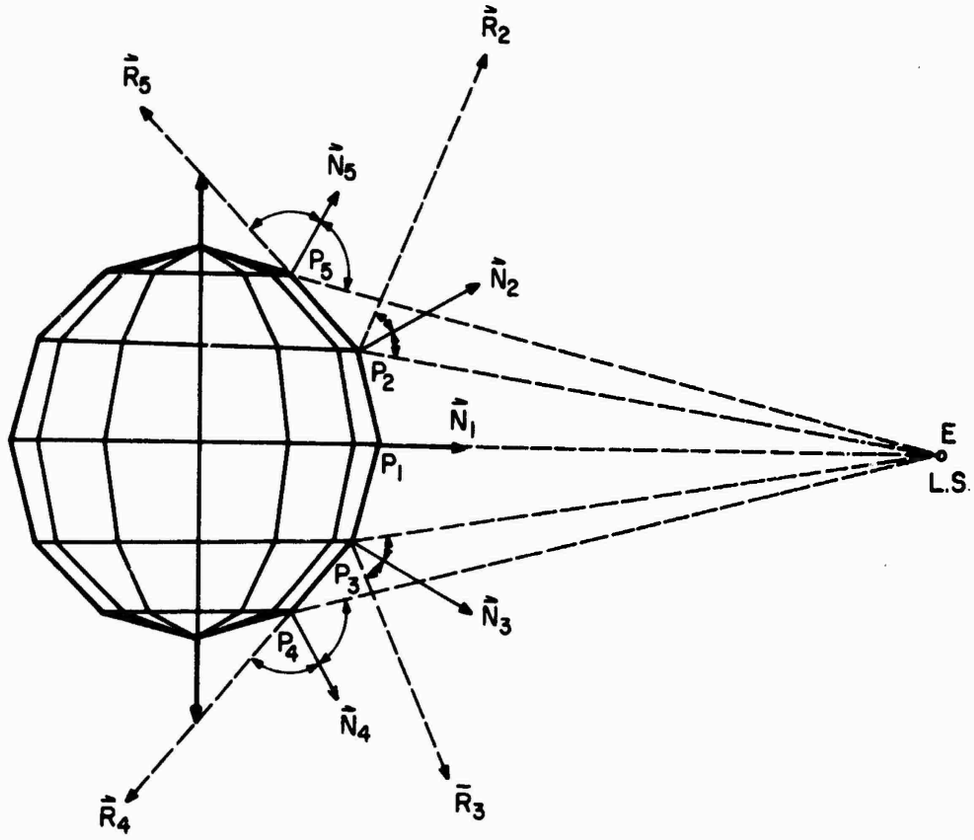
Figure 1.9 Sphere approximated by facets

circle as expected. Rather it depends on the shape of the polygons used to approximate the sphere. In the present case, the shape of the contour is a cross.

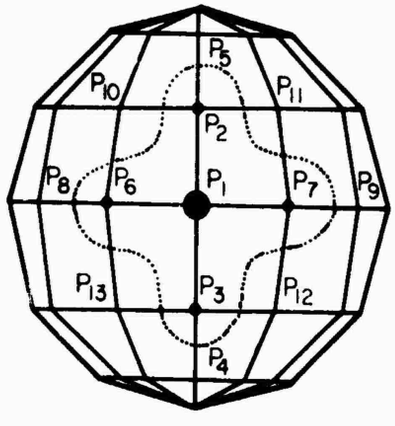
Figure 1.9 represents the line drawing of the sphere displayed on the Figure 1.8a with the cross shaped specular reflection of the light. The same sphere is shown on the Figure 1.8b with the light source oriented at 45 degrees to the line of sight. The shape of the contour changes but it is not right. This distortion in displaying a complex surface is particularly disconcerting when no information about its shape is available to the observer. In searching for an improved model, the author attempted to provide the addition of simulated specular reflection without introducing shading distortions. Techniques based on the Gouraud scheme evidenced severely shading distortions, which lead the observer to a wrong interpretation of the shape of the modeled object. A complete solution to the problem remains to be found, but is probably not possible using the piecewise model.

Finally , the problem of frame to frame discontinuities of shade in a computer generated film is illustrated in the following situation. A curved surface is approximated with planar facets. When this surface is in motion, all the facets which are perpendicular to the direction of the light take on a uniform shade. In the next frame as the motion of the object brings these facets into a different orientation toward the light, the intensity of the shade across their surfaces varies continously from one end to the other, thus creating some artificial highlights on the surface of the object. However, the position of these highlights is not steady from frame to frame as the object rotates.

A good example of this phenomenon is a cylinder turning about its axis. The cylinder is approximated with 3 planar polygons (as a prism). When one edge



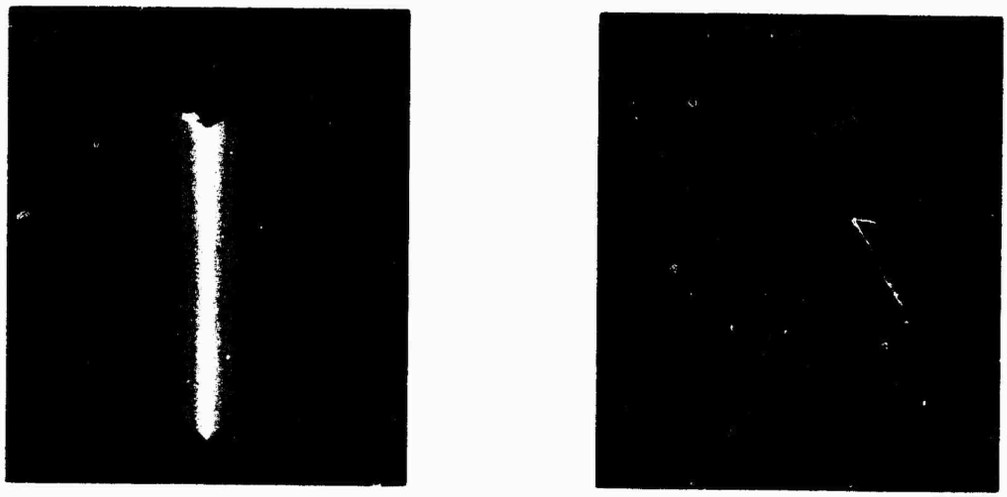
(a)



(b)

Figure 1.10

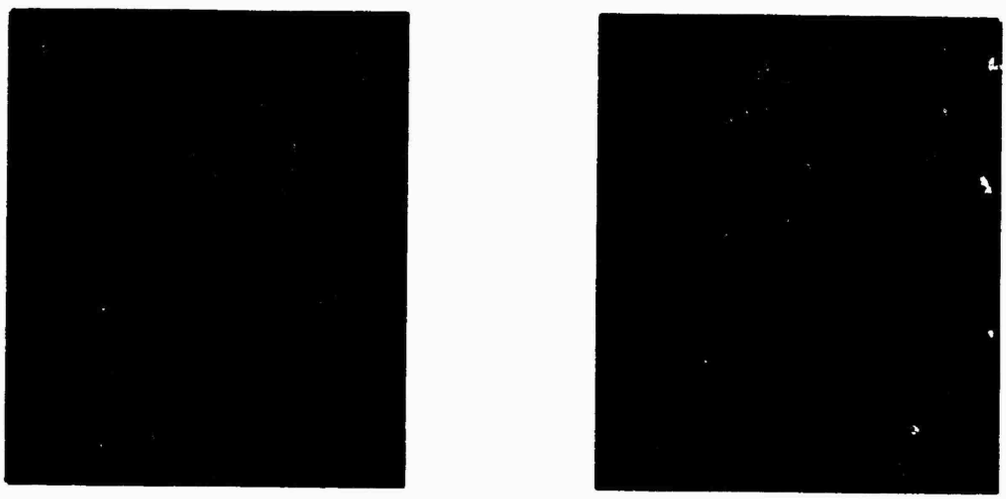
Cross-section Size of dots proportional to highlight intensity



(a) (b)

Rotation of a 3-sided cylinder model

Gouraud shading



(c) (d)

Figure 1.11

is facing the direction of the light (see Figure 1.11a) the highlight is along the edge, and the shade changes continously on both sides of the edge. In turning, the highlight position follows the edge until one face of the prism becomes perpendicular to the direction of the light. In this position, there is no more highlight on the cylinder, because the shade is uniform all over the polygon. However, the intensity of shade on the polygon is very low as shown of the Figure 1.11b. In the next frame, the highlight starts to reappear on the other side of the face as shown on the Figure 1.11c. This discontinuity becomes distracting when some of the polygons just disappear during one frame and show up in the next one.

Any research in new models and algorithms for producing shading of computer-generated images must take into account the physics and psychophysics of visual perception, as well as physical laws of optics and their problems. Therefore the next two sections are devoted to discussing these two problems, respectively. Compromise solutions are proposed to apply the theory to the restricted world of computer graphics. The fourth and fifth sections describe a method which overcomes the shortcomings of the present shading techniques just discussed. The new shading model is simple enough so that a hardware implementation can be accomplished for a real-time display system.

The appendices give information on the experiments to determine the specular reflection parameter for a set of sample materials . In addition, methods to correct for the non linearity of the film in the computer-generated pictures are presented.

CHAPTER II

VISUAL PERCEPTION

In the context of shaded picture creation, it is important to understand the fundamental properties of the human visual system. Unlike a photograph taken of the real world, a computer-generated shaded picture is made from a numerical model, which is stored in the computer as an objective description. When an image is then generated from the model of an object, the human visual system makes the final subjective analysis. Obtaining a close image correspondence to the eye's subjective interpretation of the real object is then the goal. The computer system can be compared to an artist who paints an object from its description and not from the vision of the object. But unlike the artist, who can correct the painting if it does not look right to him, the computer which generates the picture does not receive feedback about the quality of the synthetic images , because the human visual system is the final receptor. Therefore, the psychological and physiological sides of human vision must be taken into consideration, and also, the close relationship between the physical representation of the image and the reaction of the eye to this kind of representation must be studied and formalized.

The present section is devoted to giving a background on the visual perception and its problems. Although the subject has been widely studied by a large number of scientists, among them Cornsweet [21], Bekesy [22], and

Hartline [23], the author attempts to present here some problems and phenomena which are directly related to the field of computer-generated images, such as optical illusion, which is a critical part of the system.

II. 1 Representation of images

The close relationship between the information carried by the physical image and the interpretation of this information by the human vision system leads to the question of how a physical image is represented in nature. This question influences the rest of the system, therefore a choice of a particular type of representation is very important. In photography, the usual information obtained from a image is spatial variations in light intensity. The interpretation of this type of information is subjective, and therefore difficult to process. Stockham [24] showed that the light intensity I(x,y) (transmitted or reflected) of the image illuminated by an uniform light intensity io is represented by

$$I(x,y)=io*exp(-k*d(x,y)) \qquad (2.1)$$

where k is a constant, and d(x,y) is related to the density. Also, (2.2) can be written as

$$ln(I(x,y)/io)=-k*d(x,y) \qquad (2.2)$$

The notion of density is well known in photography, and as d(x,y) is proportional to the density, it is reasonable to call any logarithmic representation of an image a density representation (Stockham , 1972).

Another type of information carried by a physical image is the light energy, reflected or transmitted respectively in the case of a positive image or a

transparency. As stated by Stockham, such a representation by light intensity analogy is a relatively new practice in image technology.

At the final step of the system, a subjective interpretation of the generated image is done; it is then necessary to attempt to define the subjective brightness with respect to the density of image. The notion of subjective brightness will be discussed in the Appendix I on "Compensation Table."

II. 2 A model for human vision

A large set of psychophysical observations on experiments done with simple animal eyes have suggested that the relationship between the light intensity input to the visual receptors and the neural output level is approximatively logarithmic. Physiological evidence to support this supposition was presented by Fuortes in 1959 and Rushton in 1961 (For more detailed discussion or treatment see Cornsweet). Similar experiments with human beings have been performed and the results found were close to a logarithmic sensitivity.

The logarithmic model explains the large range of sensitivity of the eye and its capability of adaptation to the environment lighting. Also several experiments done in the determination of the just-perceptible differences of intensities between two stimuli have shown that, at central range of intensities, the eye tends to indicate a constant change when the stimulus is changed in a constant ratio. In other words, the eye is sensitive to percentage rather than to absolute changes in intensities. The eye is usually referred to as a "zero detector," which means that, when a two-part field of different intensities is presented to the eye, it can tell if they are both of same intensity or not. In the case where the two parts have different intensities, the eye can determine the

brighter one. But when each part of the field is shown separately in time and space, the eye is unable to make an accurate comparison of the shade between them.

This model of the eye will explain the notion of subjective brightness as presented in Appendix I.

II. 3 Optical Illusion

A particularly interesting aspect of visual perception is the complexity of brightness perception. In particular, some optical illusions show that the subjectivity of the visual system plays an important role in the perception of brightness. The most commonly known optical illusions are the "simultaneous contrast" and the Mach Band effect [Cornsweet, Ratliff (24)]. Simultaneous contrast may be described as the effect which makes a surface or a light source of any kind look lighter (or brighter) when it is surrounded by a dark area than when it is surrounded by a lighter one. This effect can be easilly observed in Figure 2.1. The two grey squares are of exactly same intensity, but the one surrounded by a black background is seen as the brighter of the two. This optical illusion can be explained by the spatial interaction of the brain's neural network (Cornsweet). Qualitatively, the eye tends to average the intensity of different areas of the image and seek for a common level. In the preceding example, the whole image can be seen as separated into two regions: one with black background and one with white background. In the first region, the common level is darker than in the second level. Therefore, the same grey square is seen to be brighter when it is compared to the first common level and darker when it is compared to the second common level.

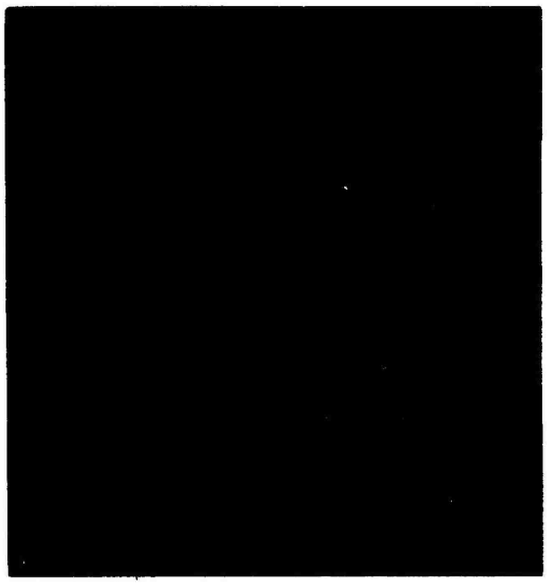


Figure 2.1 Simultaneous contrast

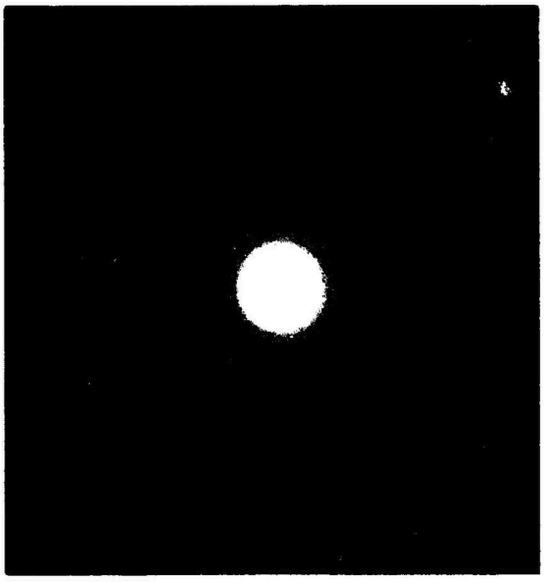


Figure 2.2 Mach Bands Effect

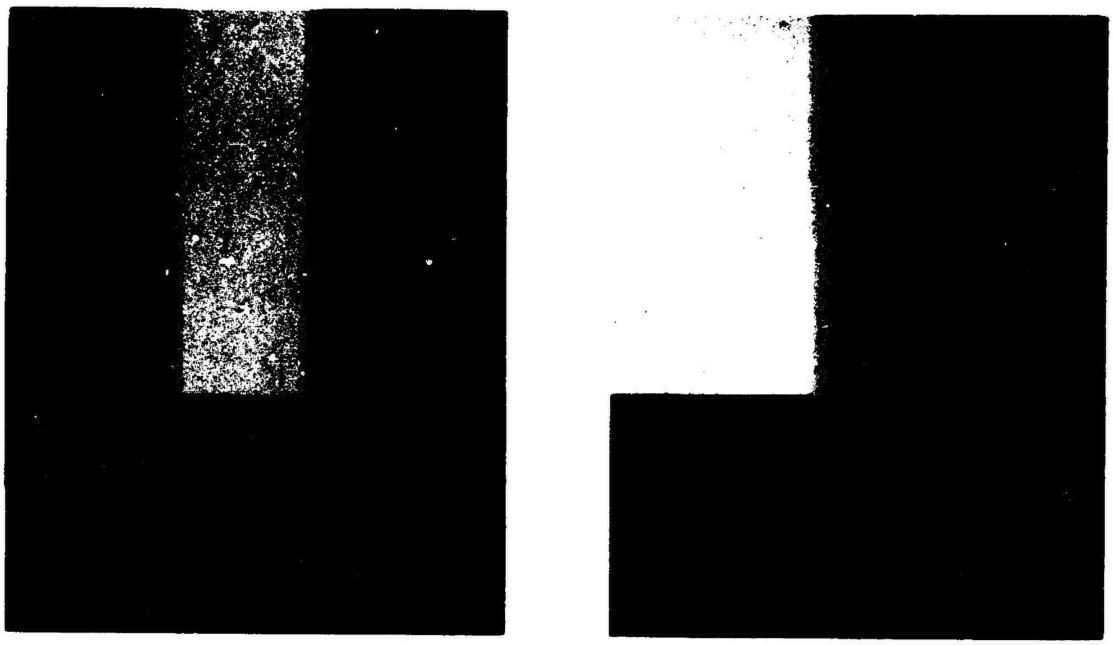
The simultaneous constrast illusion is a common phenomenon in the real world, but it is not as critical to the technique of computer-generated pictures as the Mach Band effect.

Mach established the following principle:

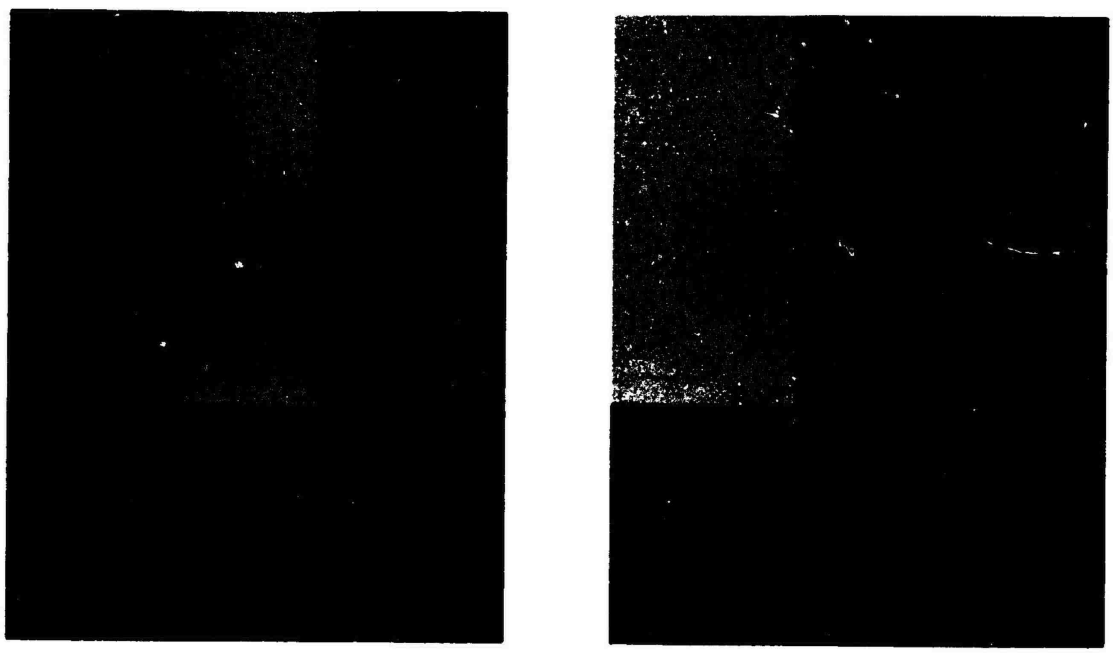
> "Wherever the light-intensity curve of an illuminated surface (the light intensity of which varies in only one direction) has a concave or convex flection with respect to the axis of the abscissa, that particular place appears brighter or darker, respectively, than its surroundings." (E. Mach, 1865);

Whenever the slope of the light intensity curve changes, this effect appears. The degree of this effect is more or less visible, depending upon the magnitude of the curvature change, but the effect itself is always present. The set of Figures 2.2, and 2.3, illustrate different Mach Band patterns.

The conclusions derived from this principle are central to the technique of displaying curved surface approximated by planar facets. The shading model which displays each facet with the same shade is suited to display objects made of planar facets, like cubes, etc., but fails to approximate a curved surface. If the goal of this technique is an attempt to restore the appearance of a smooth surface, the result is contrary to what was expected, because the eye enhances the discontinuities over polygon edges, creating undesired apparent brightnesses along the edges. By increasing the number of polygons, it was hoped that a better approximation of the curved surfaces could be obtained, and the smoothness of the approximated surface could be restored. Unfortunately, unless the size of the displayed facets is shrunk to a resolution point, the multiplication of the number of facets does not solve the problem. Using the Gouraud method to linearly interpolate the shade between vertices, the discontinuities of the shading

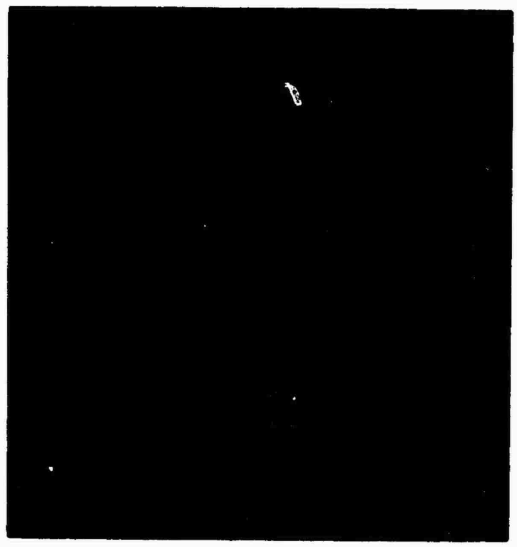


Piecewise linear curve

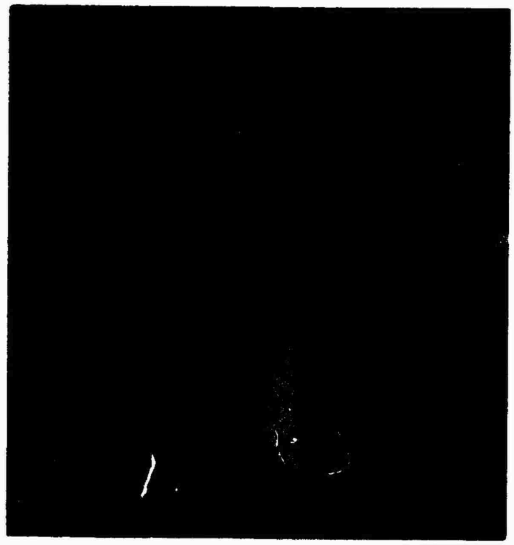


Curves having a continuous first derivative

Figure 2.3 Mach Band Effects

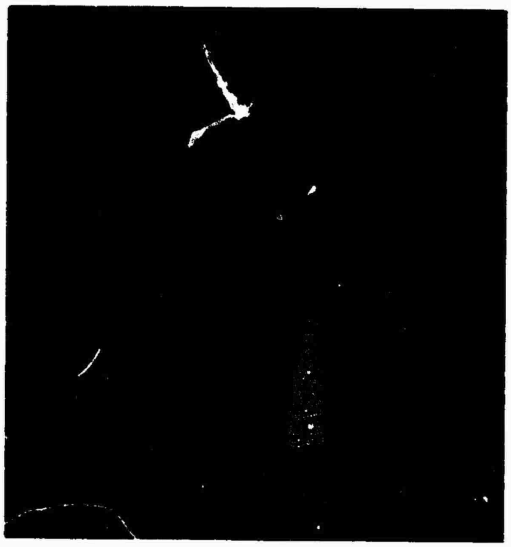


(a) 8-sided model



(b) 16-sided model

(c) 32-sided model



(d) 64-sided model

Figure 2.4 Mach Band Effect decreases with number of facets

function disappear, but the Mach Band effect is visible whithin the area where the slope of the shading function changes. The subjective discontinuity of shade at the edges due to the Mach Band effect then destroys the smooth appearance of the curved surface. The set of pictures on Figure 2.4 shows a cylinder approximated by an increasing number of polygons. The cylinder on the first picture is approximated with eight planar polygons. The distortion of the shade along the edges of the polygons is due to the Mach Band effect. In the next picture, a cylinder is approximated with sixteen planar polygons. The change of the slope of the shading function in that case is smaller, the Mach Band effect is then less visible. In order to decrease the Mach Band effect, a large number of polygons is required.

As stated by Mach himself in his papers, the Mach Band effects can be observed everywhere in the real world. In the domain of computer-generated images where a curved surface is approximated with planar polygons, the goal is to be able to decrease the effect along the edges of the polygons where apparently the shading function changes, so that the presence of the separating edges is less noticeable.

CHAPTER III

SOME PHYSICAL ASPECTS OF LIGHT AND SHADE

Before we can simulate the shade and color of any object, we must understand the true physical laws which determine these aspects as they occur in nature. Physically, light is an electromagnetic form of energy. It permeates space without attenuation and, like many other forms of energy, cannot be known to exist at all until it is converted into some other form when it strikes an object in its path or is changed in some other way. We see light only because of its effects on our eyes, and we see objects only because of the effect that they have on the light before it reaches us.

When light falls on any material, three things can occur: (1) the energy is absorbed, (2) reflected, or (3) transmitted. Usually all three occur, but the proportion of each is different at each wavelength. These three phenomena are interconnected and the effect of each of them on the visual impression of objects depends on the quality of the light and the material of which the objects are made. For the restricted world of solid opaque objects, the transmittance of light is not considered. The important question is how the color of an object can be seen.

III. 1 Physical aspects of color vision.

The color of an object is due to the combination of the absorption and the reflection of the incident light. When a beam of light falls on a material, a part of it penetrates the medium and has its energy converted to heat, with the other part reflected back. The first phenomenon is called absorption and the latter reflection. The color of the object depends on the characteristics of the reflected rays, which result from the incident light modified by the absorption characteristics of the illuminated materials.

The absorptance of a substance varies with the wavelengths of the incident light. A substance is said to show general absorption if it reduces the intensity of all wavelengths of light by nearly the same amount. Such substances appear to be grey when they are illuminated by an incident white light. However the colors of most objects are due to selective absoption; that is, absorption of certain wavelengths of light in preference to others. Practically all colored substances owe their color to the selective absorption in some part or parts of the visible spectrum. The net result of this absorption is that the light that leaves the material has a different energy distribution from the light that falls on it.

The reflection of light can be categorized into two kinds:

1. Reflection from the outer surface.
2. Reflection after the light has penetrated the material.

In the first case, except for some polished metals and other rare surfaces such as some solid dyes which have selective reflection, light reflected from the outer surface has the same quality as the illumination. This phenomenon

is known as the specular reflection of light. In the second case, some incident rays penetrate farther into the medium, and the material selectively absorbs certain wavelengths of these penetrating rays and reflects the rest in different directions. The reflected light then shows the color of the object. This phenomenon is also referred to as an internal reflection, by opposition to the external reflection by the outer surfaces of the materials.

The law of reflection states that the reflection angle r is always equal to the angle of incidence i, regardless of the nature of the material or the wavelength of the light (Figure 3.1). The only requirement is that the area under consideration be flat (Jenkins & White [26]). This very simple law gives rise to a number of possibilities, which may be summarized as follows:

1. Reflected light from a large smooth surface acts and looks as though it had come through the surface from behind.

2. If the surface is irregular two cases arise: the light may be thrown into a definite pattern or it may be diffused in all directions.

To illustrate these phenomena, polar diagrams of light reflected from surfaces with various degrees of roughness are given in Figure 3.2 (Evans [27]). The first diagram shows what is known as a "complete diffusor". This is also called a "Lambertian reflector" which was first defined for black body radiation. Such a surface appears equally bright from all directions. Each of the dashed semicircles in Figure 3.2a represents the constant brightness from different viewing angles and corresponding to an incident light. The distribution of the reflected light at different incident angles from a small area of the surface is a tangent sphere. This distribution is represented by a tangent circle in a plane diagram.

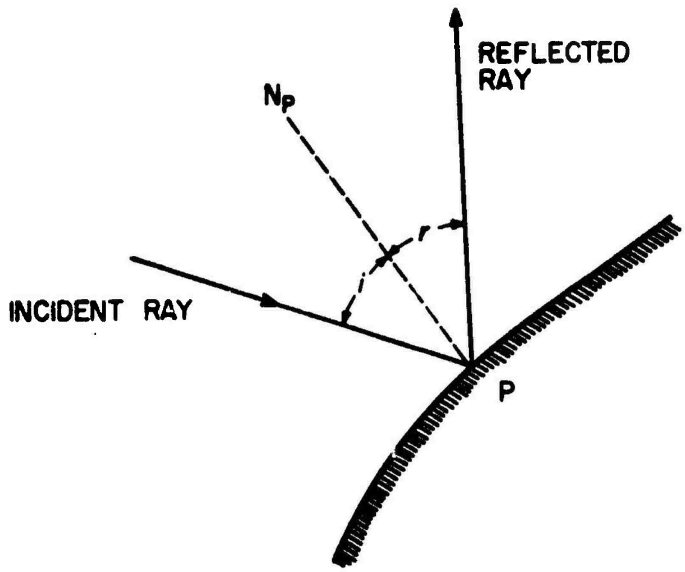


Figure 3.1 Reflection law

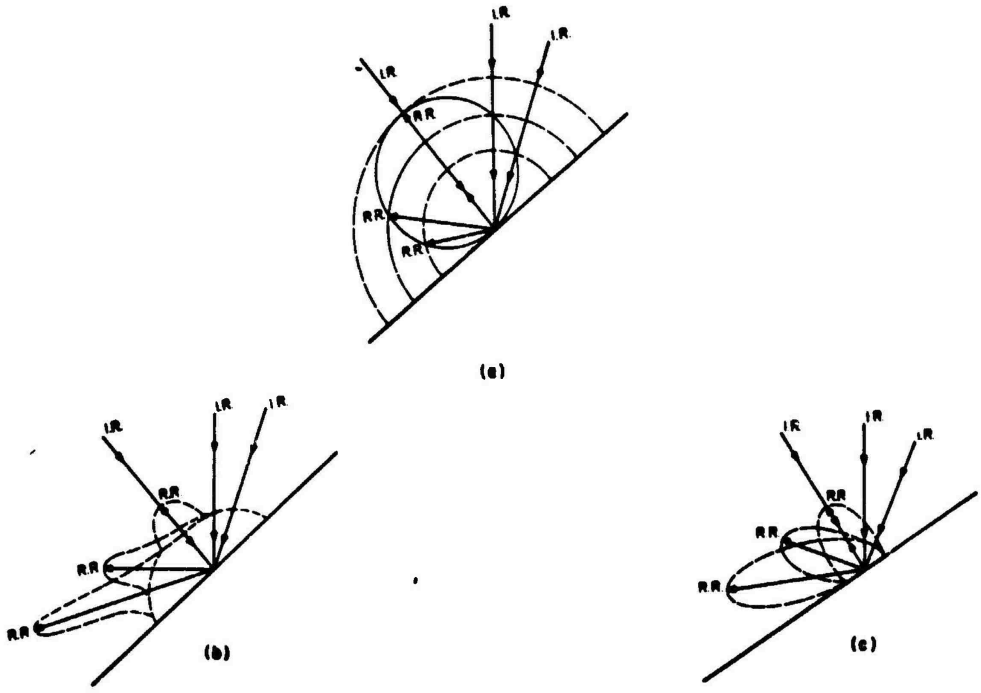


Figure 3.2 Polar diagrams of reflected light

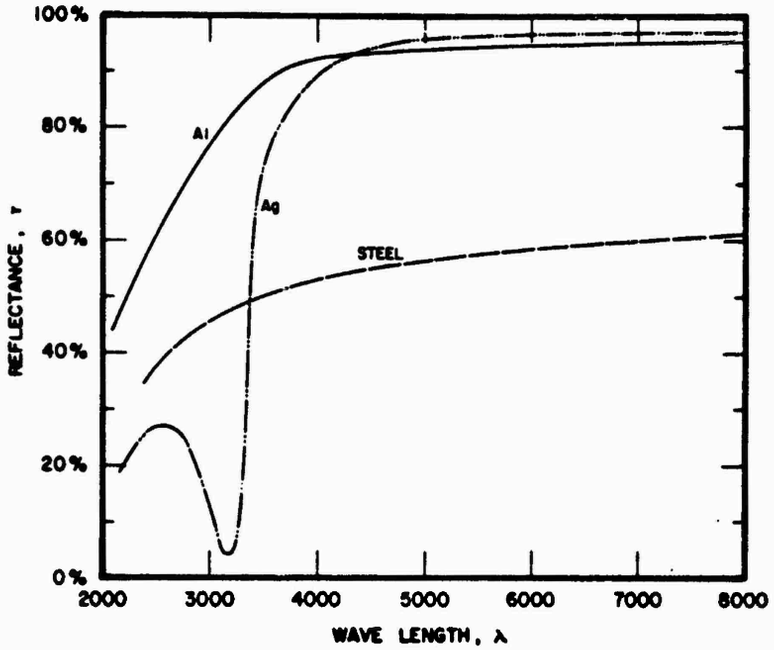
Diagrams (b) and (c) show different degrees of specular reflection. In these instances only a certain percentage of the light is said to be diffused; the remainder is specularly reflected. As the specular reflection of light is a directional phenomenon, the brightness intensity depends on the relative position of the observer to the light source and to the surface. The dashed curves represent plane diagrams of brightness intensity for a certain incident angle. The specular reflection of light is a well known phenomenon in optics and it follows the Fresnel Reflection law.

III. 2 Fresnel Law of Reflection

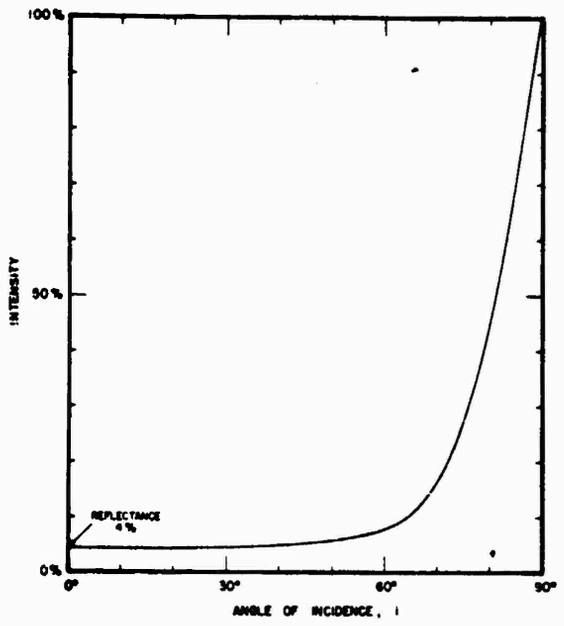
If the light strikes the surface at normal incidence, only a small percentage of the intensity of a beam of light is reflected . For example, in the case of some dielectric materials only 4 per cent of the incident light is reflected, but metals reflect 80 per cent or more of the incident light under the same conditions . At other angles of incidence the reflecting power increases with angle, at first slowly and then more rapidly until 90 degrees, the "grazing incidence" at which all of the light is reflected. The reflection curve, which is defined as the ratio of the reflected light intensity over the incident light intensity plotted against the incident angle, depends both on the nature of the object and the light source. An example of a reflection curve is given on Figure 3.3 . For more detailed discussions, see Jenkins & White [26].

III. 3 Practical Applications

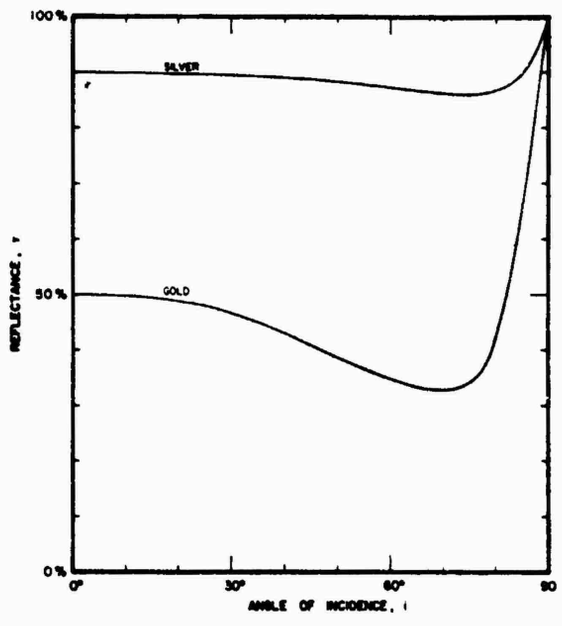
Let us take an example of a beam of light falling on a piece of material of area ds, with an incident angle i. Let the incident energy ei be the energy



(a) Reflectance at normal incidence



(b) Reflectance for a dielectric n=1.5



(c) Reflectance for some metals

Figure 3.3 Reflection curves

per unit area perpendicular to the incident ray. The energy received by the area ds is:

$$Ei = ei * ds * \cos(i) \qquad (3.1)$$

Knowing the reflection law, if Ns is the normal to the surface ds, the reflected light vector R makes an angle r=i with Ns.

The energy reflected in the direction of R is:

$$Er = Ri * ei * ds * \cos(i) \qquad (3.2)$$

where Ri is the reflection coefficient of the material relative to a wavelength. This coefficient depends also on the incident angle i.

The visual perception of the color and shade of an object depends on the amount of reflected energy received by the eye. In a restricted world where only one point source of light or one direction of light is considered, the lightness of an object depends on two phenomena:

1. Direct illumination due to the incident light.
2. Diffuse illumination due to the environment.

The direct illumination by the incident light yields two results: a specular reflection and a diffuse reflection.

The specular reflection is the reflection of the incident rays from the outer surfaces of the body of the object. The nature of the reflected rays is the same as that of the incident light. The intensity of these rays depends on the incident angle and follows the Fresnel Reflection law. The diffuse reflection is defined as the reflection of the incident rays from inner parts of the body of the

object. In this case, the intensities of the reflected rays depend on the absorption coefficients of the materials. The path of the reflected rays also changes due to reflection from successive different internal layers of the material.

Assuming that: 1) The diffuse light due to the environment is constant, and depends only on the environment, and 2) That the incident light is monochromatic, the reflected energy can be simulated as a sum of three terms:

$$Er = Edr + Esr + Ed \quad (3.3)$$

where Edr, Esr, and Ed are respectively the energy due to diffuse reflection, specular reflection and environmental diffusion of light. Resulting from the direct incident illumination, the first two depend on the incident angle i, while the environmental diffuse light is a constant.

Each term of equation (3.3) can be written as:

$$Edr = Rdr*ei*\cos(i) \quad (3.4)$$

$$Esr = Rsr(i)*ei \quad (3.5)$$

$$Ed = Rdr*ed \quad (3.6)$$

where Rdr is the reflection coefficient for a material and a wavelength, Rsr(i) is the specular reflection coefficient, which is a function of the incident angle and ed is the incident diffuse energy due to the environment.

Replacing these terms in equation (3.3), it can be written as:

$$Er = Rdr*ei*\cos(i) + Rsr(i)*ei + Rdr*ed \quad (3.7)$$

As ed is a portion of ei, then the ratio d=ed/ei can be used in (3.7):

$$Er = Rdr*ei*(d+\cos(i)) + Rsr(i)*ei \qquad (3.8)$$

By analogy between the energy and the intensity of light, the reflected intensity in the direction of R can be written as:

$$Ir = Ii*Rdr*(\cos(i)+d) + Ii*Rsr(i) \qquad (3.9)$$

The first term which is a sum of diffuse reflection of light, therefore does not depend on the position of the observer, but only on the relative position of the light to the object. On the other hand, the specular reflection of light is directional and it is a function of both the positions of the light, the object and the observer. The lightness of an object can be then simulated by knowing the relative positions of the object, the light source and the eye.

CHAPTER IV

ILLUMINATION MODEL

IV. 1 Curvature of the surface at each vertex.

For a mathematically described model, the curvature of its surface at any point can be determined exactly by solving the equation which determines the surface at that point. In the case of a polygonally described model, only an approximation of its curvature can be obtained.

The curvature of a polygonally described surface is obtained from the orientation and magnitude of the normal to the surface at that point.

The determination of the curvature at each vertex of the polygons approximating a surface is then reduced to the computation of the normal to the surface at each vertex. The graph in Figure 4.1 shows a characterization of each method used to approximate the normal to the surface at a vertex. Two commonly used techniques are:

1. Approximation by averaging the normals of the surrounding polygons.
2. Approximation from the edges which terminate at the vertex.

In the first case, the normal to a polygon is the sum of the cross products of all the adjacent pairs of edges of the polygon. In the example on the diagram of Figure 4.2, the normal to the polygon P1, P2, ..., P6 is:

A CHARACTERIZATION OF DIFFERENT METHODS FOR COMPUTATING THE NORMAL AT A VERTEX

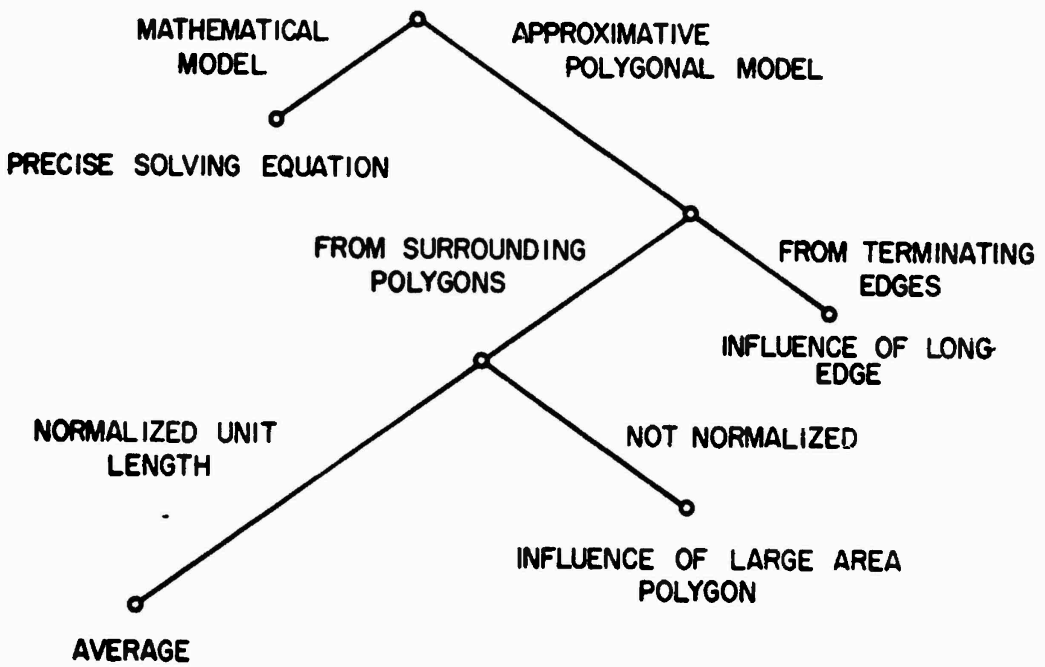


Figure 4.1 Characterization of different methods for normal computation

N=P1P2/\P2P3+P2P3/\P3P4+P3P4/\P4P5+P4P5/\P5P6+P5P6/\P6P1+P6P1/\P1P2 (4.1)

If O is the coordinate origin, the previous expression can be simplified by substituting P1P2=OP2-OP1, P2P3=OP3-OP2, etc., and we will have:

N'=OP1/\OP2+OP2/\OP3+OP3/\OP4+OP4/\OP5+OP5/\OP6+OP6/\OP1 (4.2)

with N'=k*N.

Another technique for computing the normal to a polygon is to project the polygon onto the three principal planes (Ox,Oy), (Oy,Oz), and (Oz,Ox) of the coordinate system. The projected areas on each of these planes are the three components in terms of the normal to the polygon. This method is mathematically the same as the method of cross-products, but it is more economical in computation time.

Once the normals to all the polygons surrounding a vertex are computed, they are added together and the result is assigned to this vertex. If the normals to the polygons sharing a vertex are normalized to have unit length before they are added together, the individual area of each polygon will not have any influence over the resulting direction of the normal at that vertex. The normalization can be done either by giving to each normal a unit length, or by averaging the normal to each polygon by dividing by the number of edges of each polygon. In the first case, the normal at the vertex is an average of the normal of the surrounding polygons. In the latter case, polygons with fewer edges have more influence than many-edged polygons.

If the normals to the polygons are not normalized, then because the length of a normal is proportional to the area of the polygon, large area polygons

will exert more influence on the orientation of the approximated normal at each vertex.

In the case where the normal to the surface at a vertex is approximated from the edges which terminate at the vertex, the normal is the sum of the cross-products of these edges. In the example in Figure 4.3, the normal at vertex P0 is:

$$N0=P0P1/\backslash P0P3+P0P3/\backslash P0P5+P0P5/\backslash P0P7+P0P7/\backslash P0P1 \qquad (4.3)$$

In that case, the length of an edge is the principal factor determining the orientation of the normal at the vertex.

The difference of shade in the resulting picture due to using different techniques to approximate the normal at each vertex is nearly unnoticeable. Gouraud even suggested [19] that the determination of the normal can be done by "guessing," and he showed that the result was not much different from that obtained using other techniques. Therefore, any scheme used to approximate the normal seems to be adequate for the computation of object shading.

IV. 2 Normal at a point on the surface.

The normal at each vertex can be approximated by either one of the previous methods. It is necessary now to define the normal to the surface along the edges and at a point on the surface of a polygon.

The normal to the surface at a point along the edge of a polygonal model is the result of a linear interpolation to the normals at the two vertices of that edge. An example is given in Figure 4.4: the normal Nt to the surface at a point

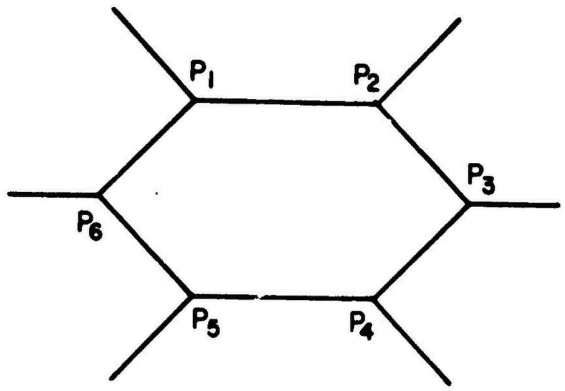


Figure 4.2 Polygon area

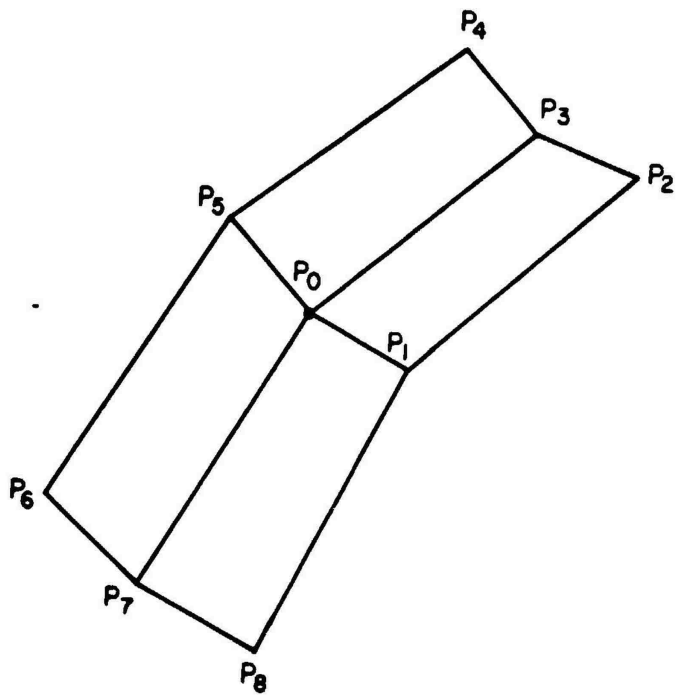


Figure 4.3 Terminating edges

between the two vertices P0 and P1 is computed as followed:

$$Nt=t*N1+(1-t)*N0 \qquad (4.4)$$

where t=0 at N0 and t=1 at N1.

For t = 1/4, N1/4 = (1/4)*N1 + (3/4)*N0.

The determination of the normal at a point on the surface of a polygon is achieved in the same way as the computation of the shading at that point with the Gouraud technique. The normal to the visible surface at a point located between two edges is the linear interpolation of the normals at the intersections of these two edges with a scan plane passing through the point under consideration. Note that the general surface normal is quadratically related to the vertex normal.

From the approximated normal at a point, a shading function will determine the shading value at that point.

IV. 3 Shading function model.

In computer graphics a shading function is defined as a function which yields the intensity value of each point on the body of an object from the characteristics of the light source, the object, and the postion of the observer.

By applying the equation (3.9) of the previous chapter, the shading at a point P (see Figure 4.5) on an object can be computed as:

$$Sp=Cp*[\ \cos(i) + d\ ] + W(i)*[\ \cos(s)\ ]\hat{}n \qquad (4.5)$$

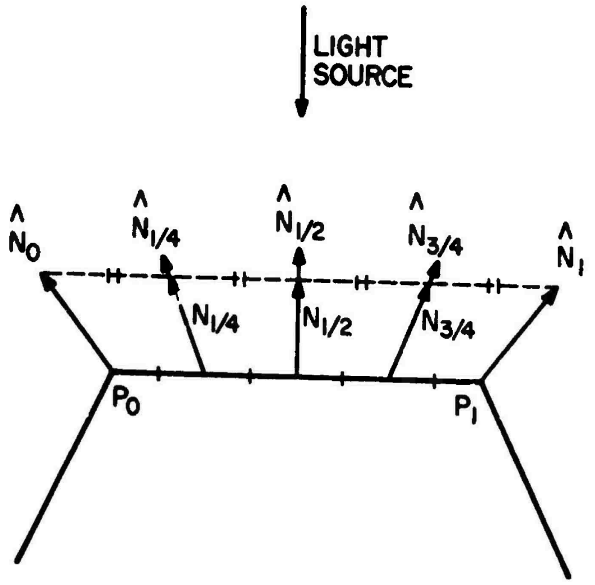


Figure 4.4 Normal at a point along an edge

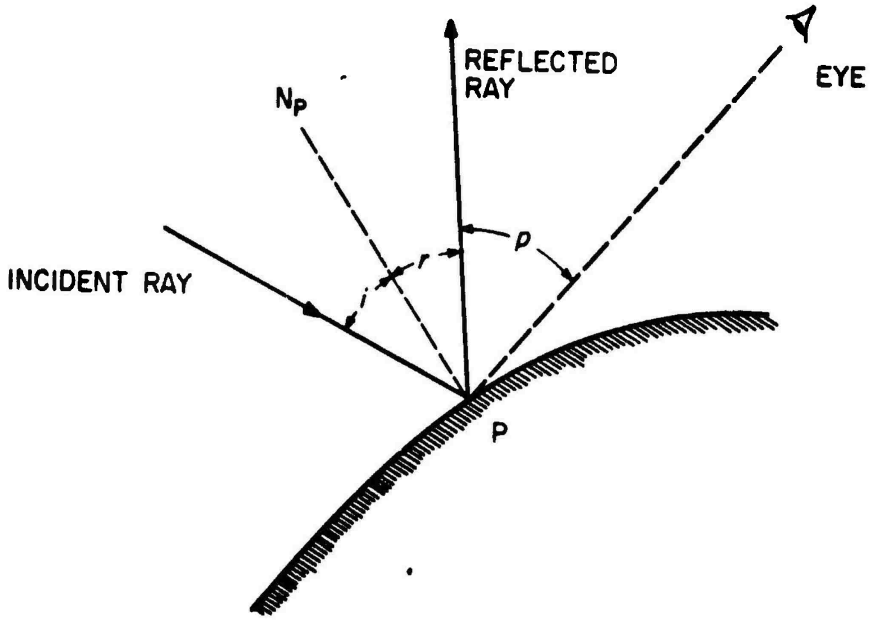


Figure 4.5 Shading at a point

Cp is the reflection coefficient of the object at point P for a certain wavelength

i is the incident angle

d is the environmental diffuse reflection coefficient

W(i) is a function which gives the ratio of the specular reflected light and the incident light as a function of the incident angle i

s is the angle between the direction of the reflected light and the line of sight

n is a power which models the specular reflected light for each material.

The function W(i) and the power n express the specular reflection characteristics of a material. Their determination for a sample of various materials and the empirical method used in this process are described in Appendix II.

In order to simplify the model, and thereby the computation of the terms cos(i) and cos(s) of formula (4.5), it is assumed that:

1. The light source is located at infinity; that is, the light rays are parallel.
2. The eye is also removed to infinity.

With these two considerations, the shading function in (4.5) can be rewritten as:

$$Sp=Cp*[\ k*Np/|Np| +d\ ] + W(i)*[\ u*Rp/|Rp|\ ]\uparrow n \quad (4.6)$$

where k and u are respectively, the unit vectors in the direction of the light and the line of sight, Np is the normal vector at P, and Rp is the reflected light

vector at P.

The quantity k*Np/|Np| can be referred to as the projection of a normalized vector Np on an axis parallel to the direction of the light. If |Np| is unity, the previous quantity is one component of the vector Np in a coordinate system where the direction of light is parallel to one axis. In this case, the quantity u*Rp/|Rp| can be obtained directly from the vector Np in the following way:

Let us consider a cartesian coordinate system having the origin located at point P and having the z axis parallel to the light but opposite in direction (Figure 4.5).

We have the following assumptions about the model:

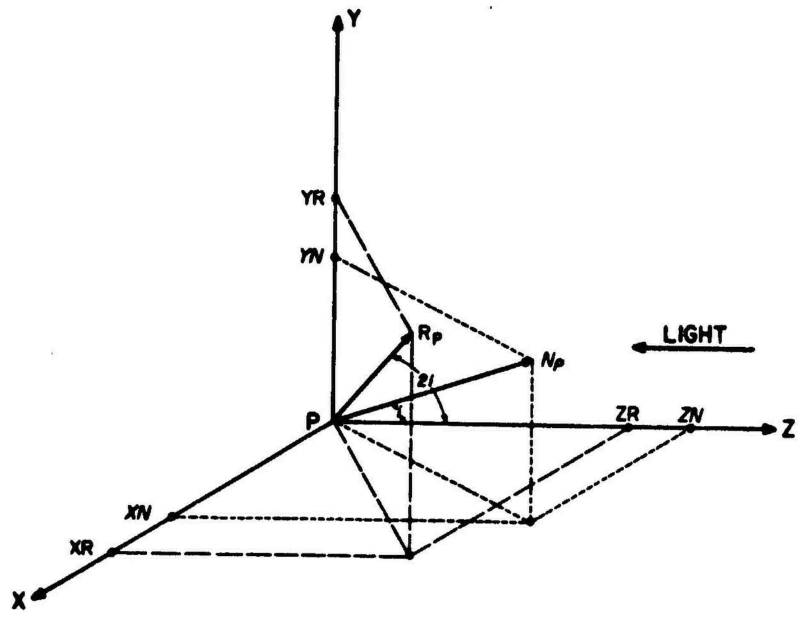
1. The normalized vector Np makes an angle i with the z axis and the reflected light vector Rp makes an angle 2*i with the same axis.

2. Only incident angles less than or equal to 90 degrees are considered in the shading computation. For a greater angle, this means that the light source is behind the front surface. In the case where a view of the back surface is desired when it is visible, it can be assumed that the normal will always point toward the light source.

3. If k is the unit vector along the Pz axis, then by simple geometry, it may be shown that the three vectors k, Np, and Rp are co-planar.

4. The two vectors Np and Rp are of unit length.

From hypothesis (3), the projections of the vectors Np and Rp onto the plane defined by (Px,Py) are merged into a line segment (Figure 4.6b). Therefore,



(a)

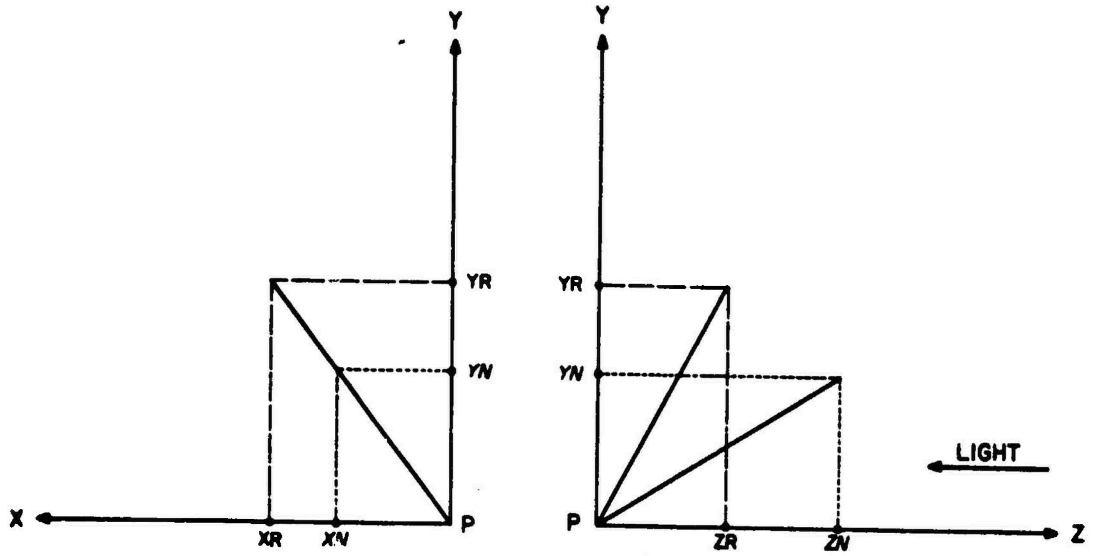


Figure 4.6 Determination of the reflected light direction

$$(Xr/Yr) = (Xn/Yn) \qquad (4.7)$$

where Xr, Xn, Yr, and Yn are respectively the components of Rp and Np in x and y directions.

From hypotheses (1) and (2), the component Zn of Np is:

$$Zn = \cos(i) \qquad (4.8)$$

where $0 <= i <= 90$ degrees.

By simple trigonometry, we obtain the following expressions:

$$Zr = \cos(2*i) = 2*[\cos(i)]\uparrow 2 - 1 = 2*Zn\uparrow 2 - 1 \qquad (4.9)$$

$$Xr\uparrow 2 + Yr\uparrow 2 = [\sin(2*i)]\uparrow 2 = 1 - [\cos(2*i)]\uparrow 2 \qquad (4.10)$$

From (4.7) and (4.10), we obtain:

$$Xr = 2*Zn*Xn$$

$$Yr = 2*Zn*Yn$$

where $0 <= Zn <= 1$.

The three components of Rp are then known in the light source coordinate system. By a simple transformation which rotates the light source coordinate system into the eye coordinate system, hence the projection of the vector Rp onto one of the axes of the new coordinate system will be known. The component of Rp on an axis parallel to the line of sight is the value of the cosine of the angle between the reflected light and the line of sight. The value of this cosine will be used in the simulation of the specular reflection of light.

This method of calculating the direction of the reflected light for each point from the orientation of the normal is preferred over the computation of the reflected light vector at vertices and the subsequent interpolation of them in the same way as the normal. This is faster and it requires less storage space than the interpolation scheme.

With the described method, the shading of a point is computed from the orientation of the approximated normal; it is not a linear interpolation of the shading values at the vertices. Therefore, a better approximation of the curvature of the surface is obtained and highlights due to the simulation of specular reflection are properly rendered.

The linear interpolation scheme used here to approximate the orientation of the normal does not guarantee a continuous first derivative of the shading function across an edge of a polygonal model. In extreme cases where there is an abrupt change in the orientation of two adjacent polygons along a common edge, the subjective brightness due to the Mach band effect will be visible along this edge. However, this effect is much less visible in the described model than in the Gouraud smooth shading model. Also, an interesting fact discussed in section II on "Visual Perception" shows that the Mach Band Effect is visible whenever there is a great change in the slope of the intensity distribution curve, even if the curve has a continuous first derivative. When a higher degree interpolation curve is used, it will make the presence of the edges unnoticeable, but will still give some Mach Band effect, as occurs in the real world.

In order to have a continuous first derivative interpolation function, cubic splines and polynomials can be used. However, as time is the critical factor in a real-time dynamic picture display system, the use of a cubic

interpolation curve does not seem to be possible at the moment with the current techniques to compute the coefficients of a cubic function. Discussions of an eventual implementation and problems involved in the use of an interpolation curve of higher degree will be presented later.

IV. 4 Comparison of pictures.

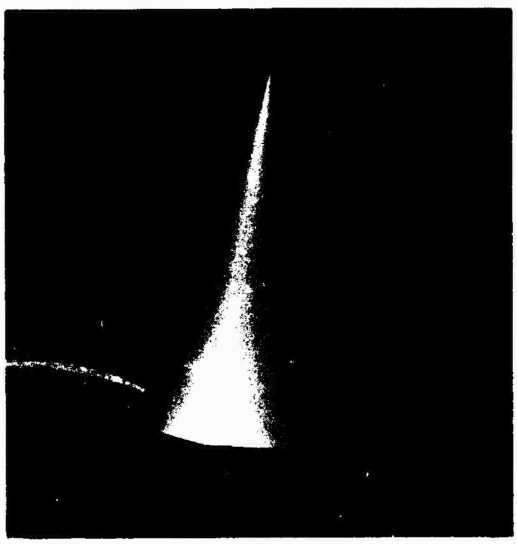
Many pictures have been generated with the described illumination model. They are first compared with pictures of objects rendered with different shading techniques, and then with pictures of real objects like cones, cylinders and spheres.

When the pictorial comparison is made between synthetic images generated with different shading techniques, the numerical model used and the lighting simulation conditions are the same for each method. In order to have a good comparison between the synthetic pictures generated with the improved shading method and pictures of real objects, a larger numerical data base is used in the model. The sole purpose of this is to smooth the silhouette of the synthetic image, so that the observer is not distracted by the non-smooth profile of the rendered objects. When considering only the shading quality of the synthetic images, a much smaller data base is needed to obtain the same result, but does not give a completely smooth contour.

The pictures of a 12-sided cone model in Figure 4.7 have two purposes: first, they give a pictorial comparison between different shading techniques, and second, the picture generated with the improved method shows that the quality of the shading is still high with a smaller data base.

(a) faceted shading

(b) Gouraud shading



(c) improved shading

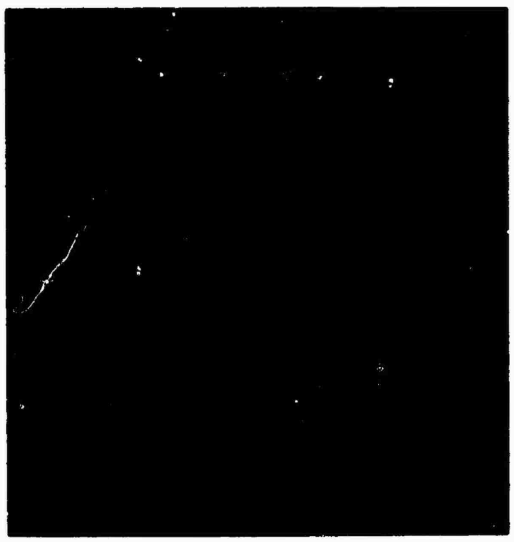
Figure 4.7 Cone rendered by different shading techniques

Pictures of a B-58 airplane are generated with different shading techniques: picture of the B-58 in Figure 4.8 is generated with Gouraud method. Apparent highlights along the polygon edges are visible. Figure 4.10a shows the same airplane displayed with the improved shading technique and without highlights simulation. When highlights are simulated in Figure 4.10b, the picture seems to have more "life."

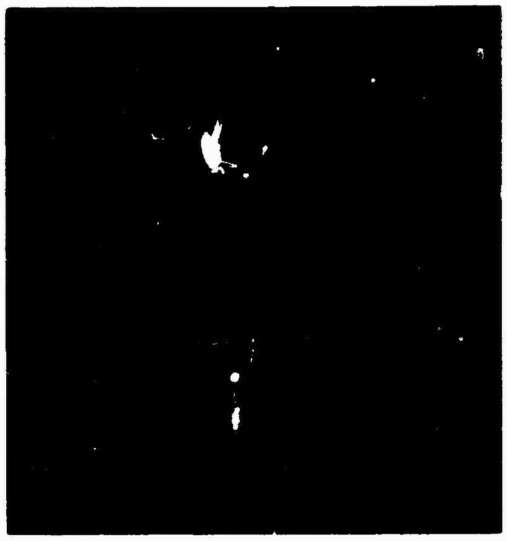
Pictures of transparent objects in Figure 4.9 are generated using the same technique as Newell et al. to simulate transparency. Highlights on the model using Gouraud shading are less intense and sometimes missing, because the Gouraud technique produces less information about the orientation of the normal at every point on the surface of a polygon. This is especially true in the case of a large polygon directly facing the light source; since the shading at each point on the surface of the polygon is a linear interpolation of the shadings at the vertices, the polygon will have a uniform shading over all of its surface, and therefore no highlights can be simulated inside an individual polygon. As an example, synthetic pictures of a transparent goblet are generated with highlights simulation using both the Gouraud and the improved shading techniques. In the picture using the improved technique, highlights can be seen on the handle of the goblet, but not in the picture generated with Gouraud shading. This is due to the size of the highlights. The Gouraud technique can form highlights only along the edges of a polygon. The improved technique is not limited in this way. This gives more information about the shape of the surface and the generated pictures look more realistic. Figure 4.14 shows the same goblet placed in front of a chess-board. In this picture a white incident light is simulated and the color of the highlights is the same as the illumination.

Gouraud shading

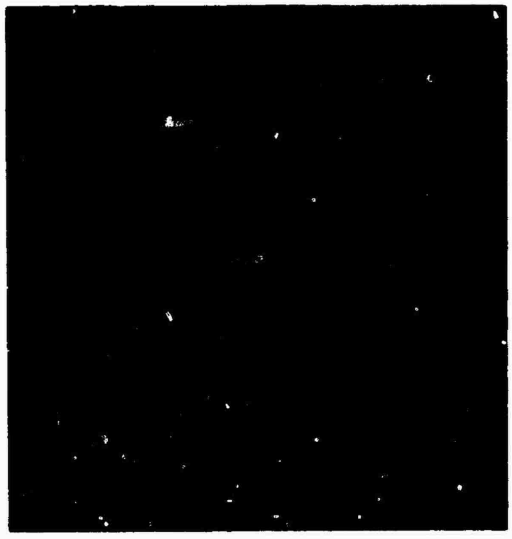
Figure 4.8 B-58 airplane



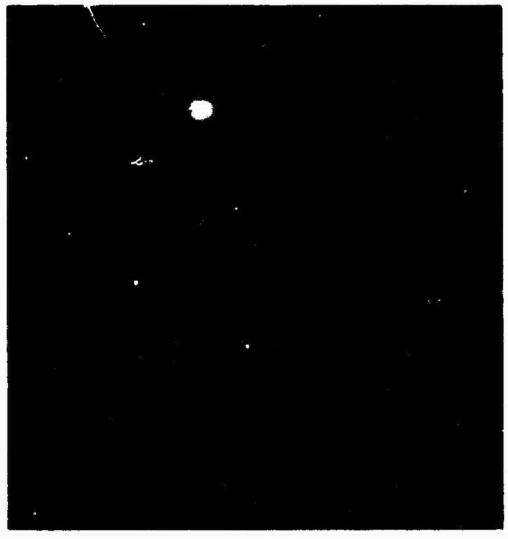
(a) specular reflection with Gouraud shading



(b) improved shading

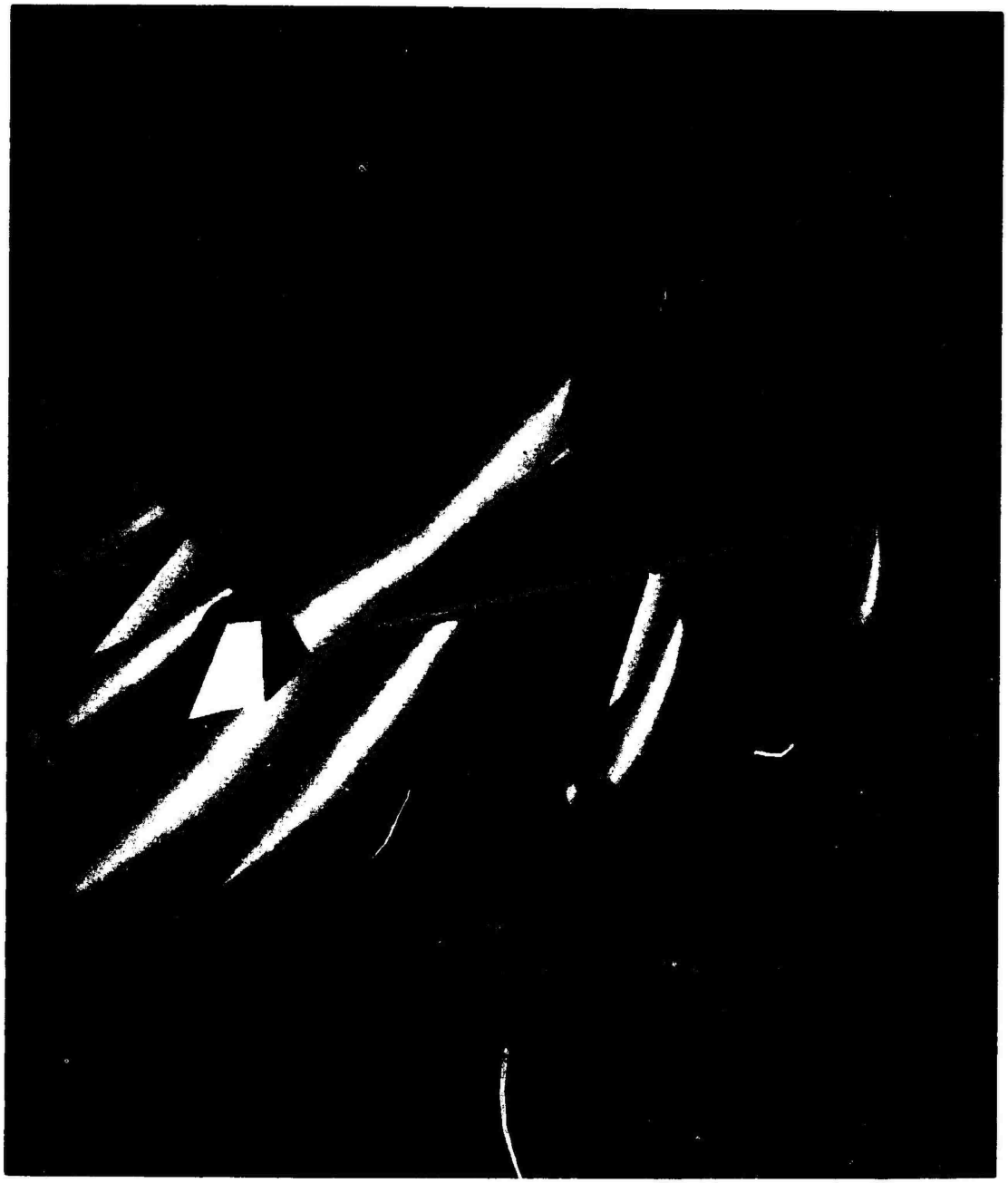


(c) specular reflection with Gouraud shading



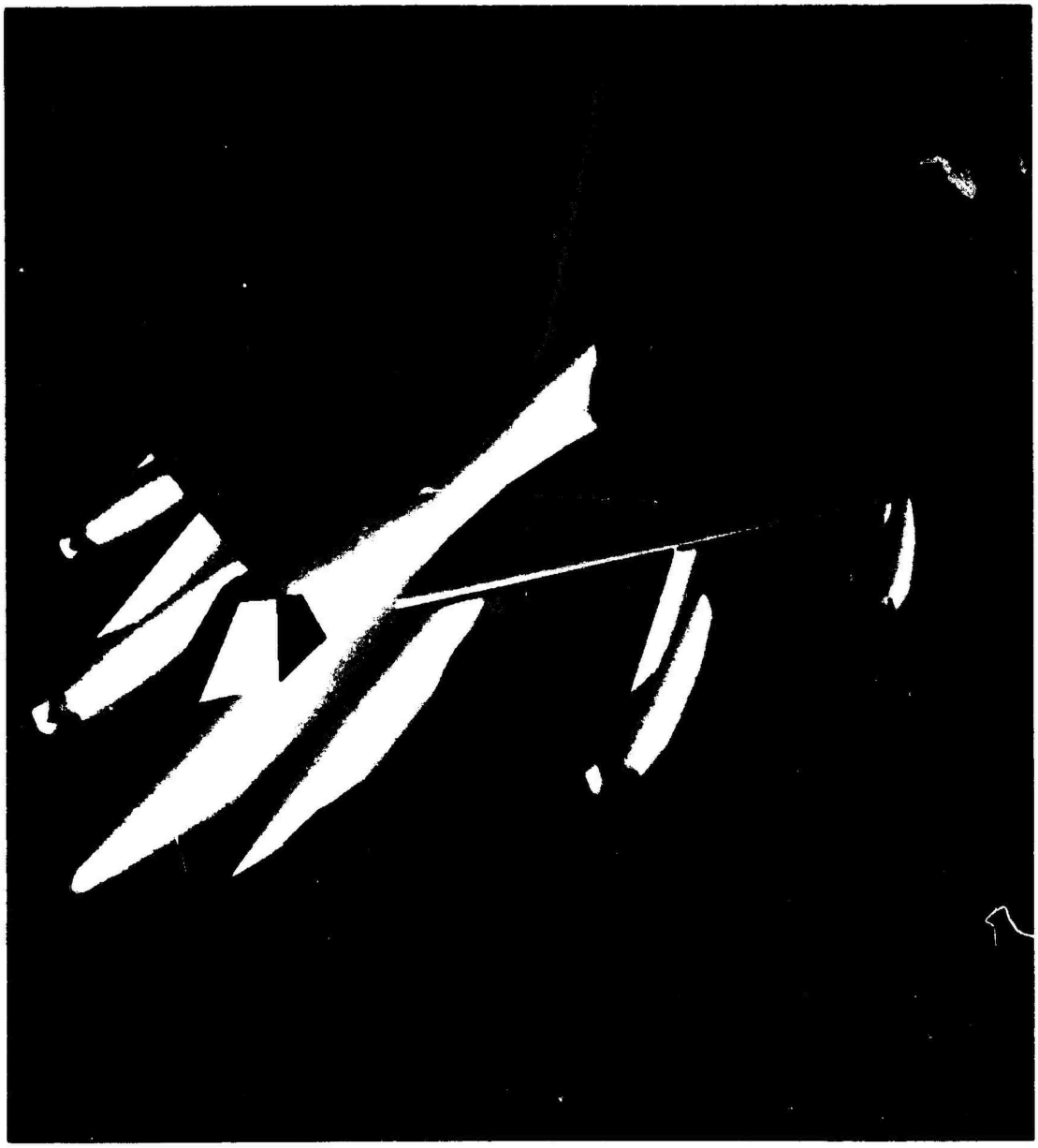
(d) improved shading

Figure 4.9 Transparent objects



Improved shading without highlights

Figure 4.10a B-58 airplane



Improved shading with highlights

Figure 4.10b B-58 airplane

Figure 4.11 shows a molecule generated with Gouraud shading and with the improved shading technique. The incident light is coming from the upper right corner.

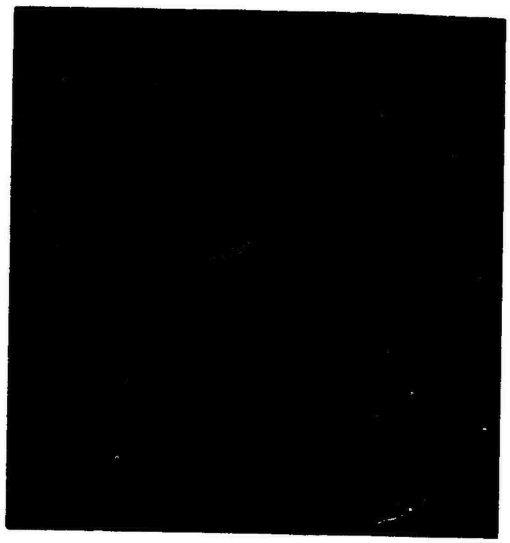
Different shading methods for polygonal models are summarized in the four pictures in Figure 4.12: a molecule model is generated with faceted, Warnock, Gouraud and improved shading methods. The addition of highlights on the pictured model gives more "life" to the object, even with Warnock's method, which is a non continuous shading method. A picture of the same model molecule generated with the described illumination technique having the light source placed at the upper right of the observer is shown in Figure 4.13.

The comparison of the described illumination model with different shading techniques demonstrates the improvement in the quality and realism of the computer-generated images. A further step is the comparison of computer-generated pictures with photographs of real objects taken under "similar" conditions.

The following pictures of real objects with simple shapes are taken in an environment where there is no diffuse light and the incident light is from a point source. The objects are made of wood and painted with different surface finishes. The comparison is left to the the observer. The author would like to mention that the exact recreation of the highly reflecting cylinder can be obtained with the improved shading technique. This may not be desirable, as no photographer ordinarily take such pictures as the ones shown in Figure 4.18b. Due to the short dynamic range of the print film, the whole cylinder can not be reproduced exactly as it would be seen by a human observer.

All the pictures generated here have been made with a constant ratio of

(a) Gouraud shading

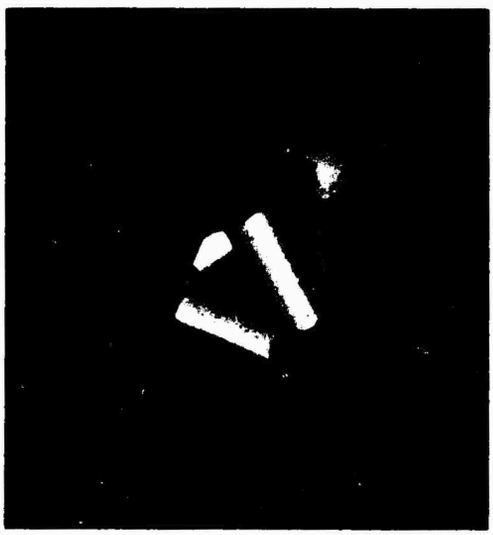


(b) improved shading

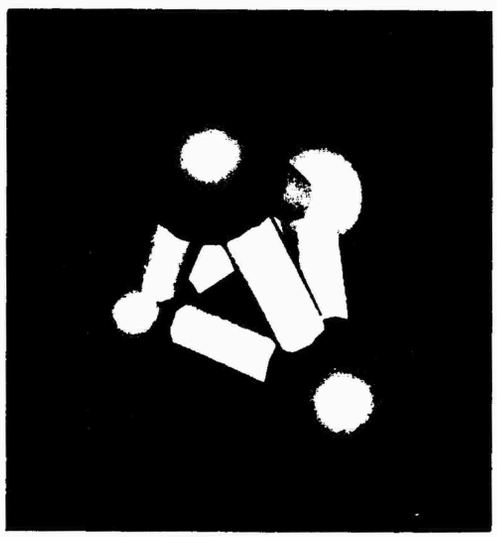
Figure 4.11 Molecule

WARNOCK SHADING

WARNOCK SHADING WITH HIGHLIGHTING



GOURAUD SHADING



PHONG IMPROVED SHADING

Figure 4.12 Molecule with different shading techniques

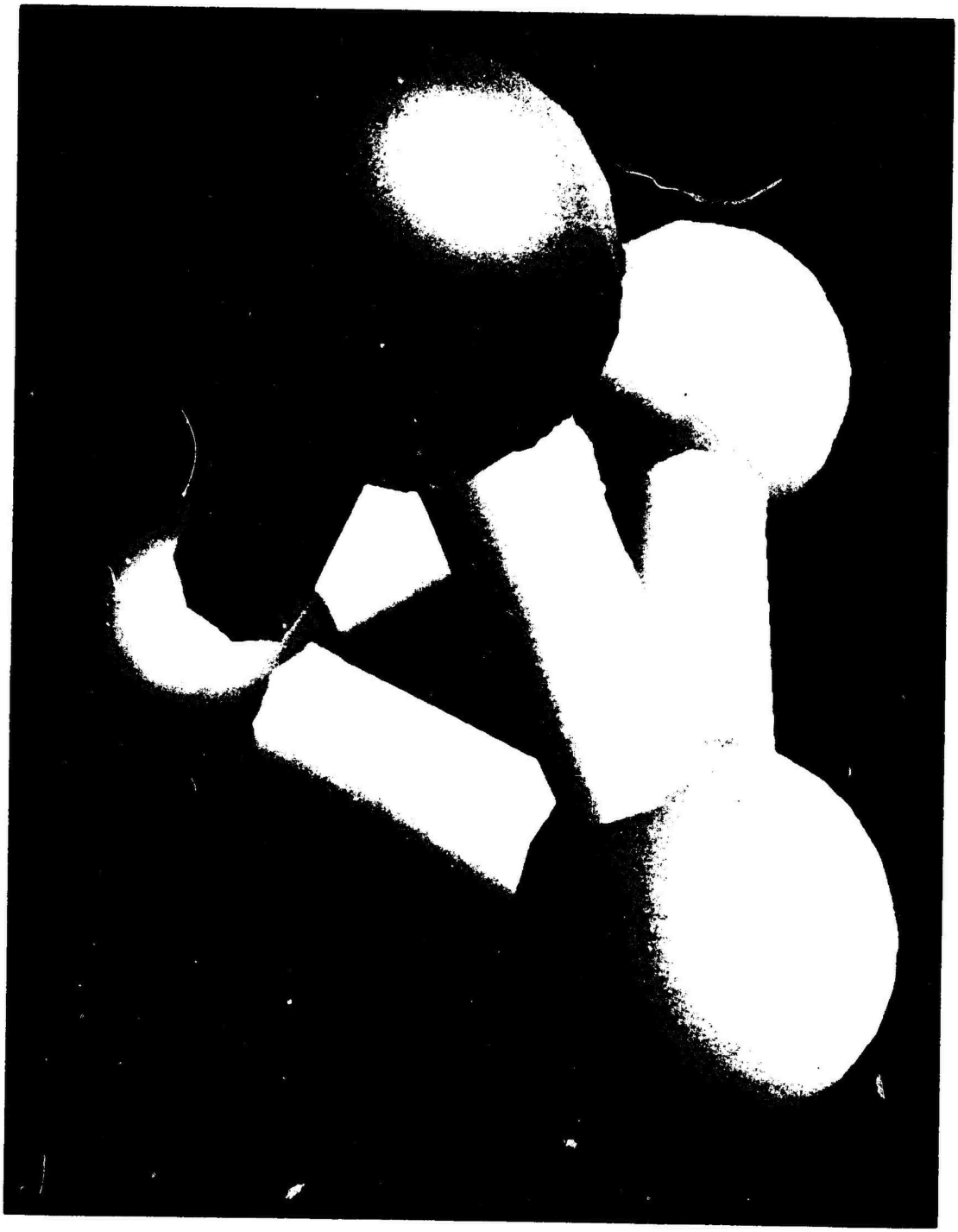


Figure 4.13 Molecule with improved shading

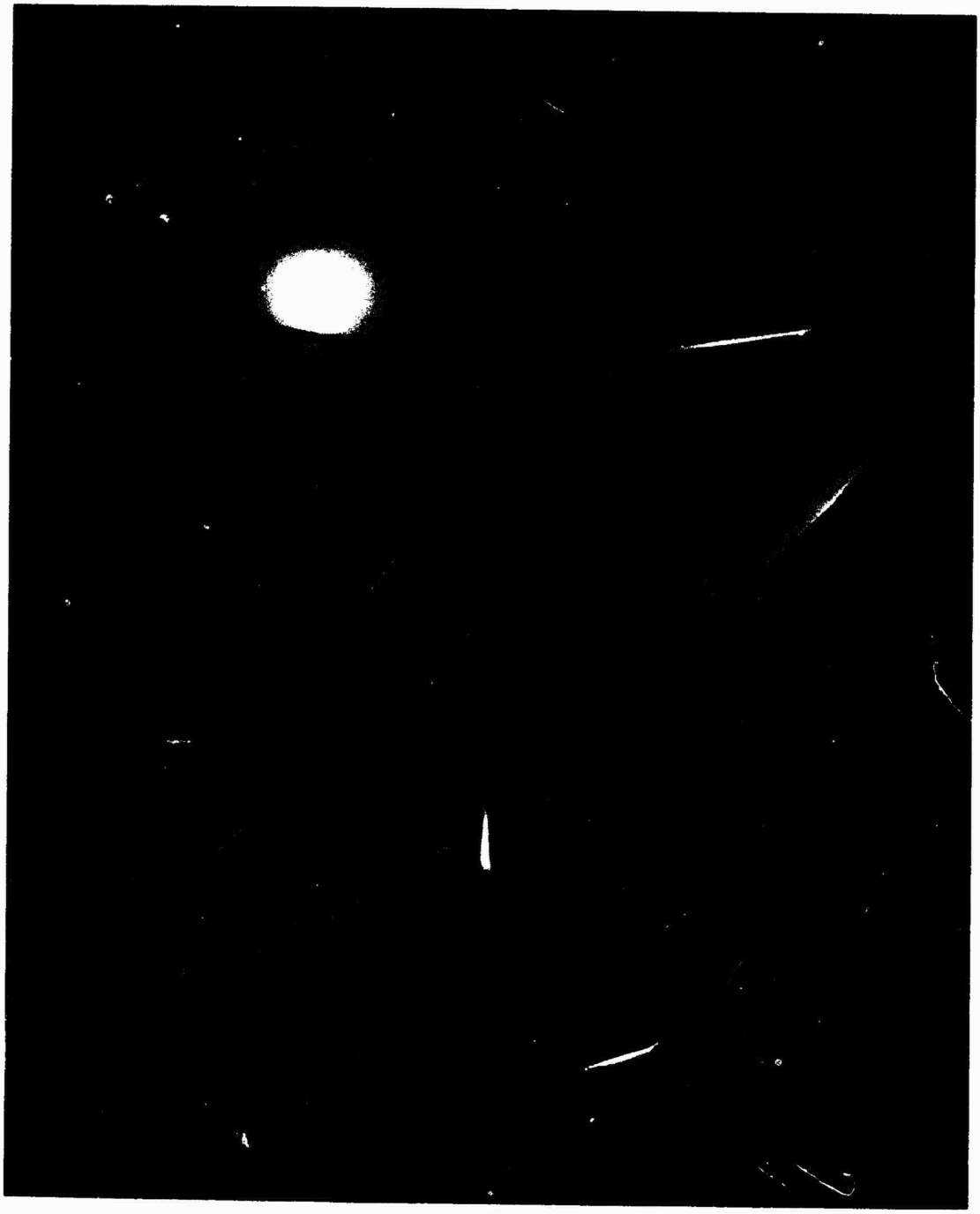
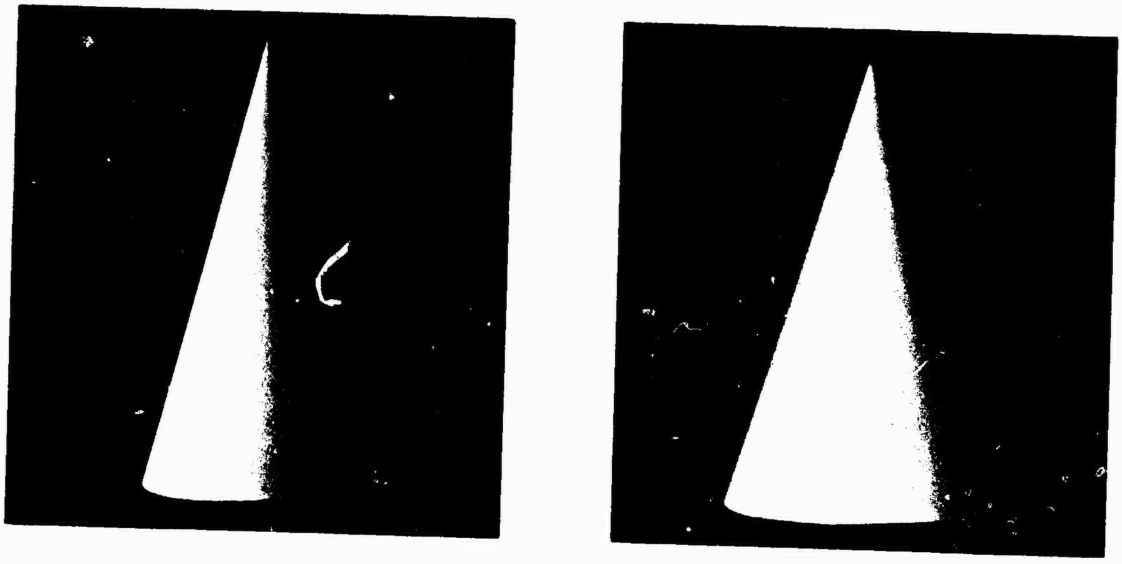


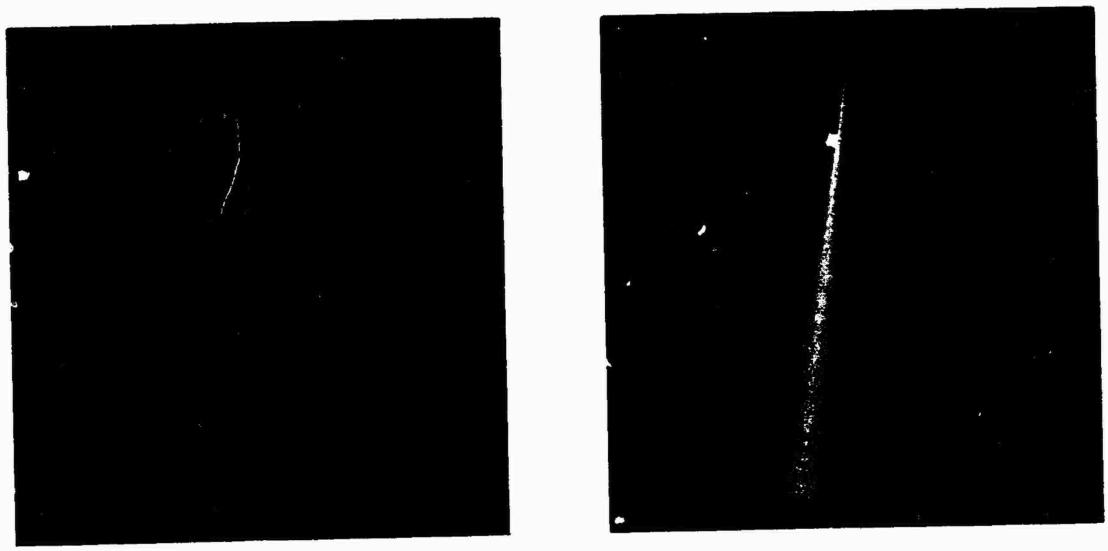
Figure 4.14 Transparent cup with improved shading

(a) aluminum paint

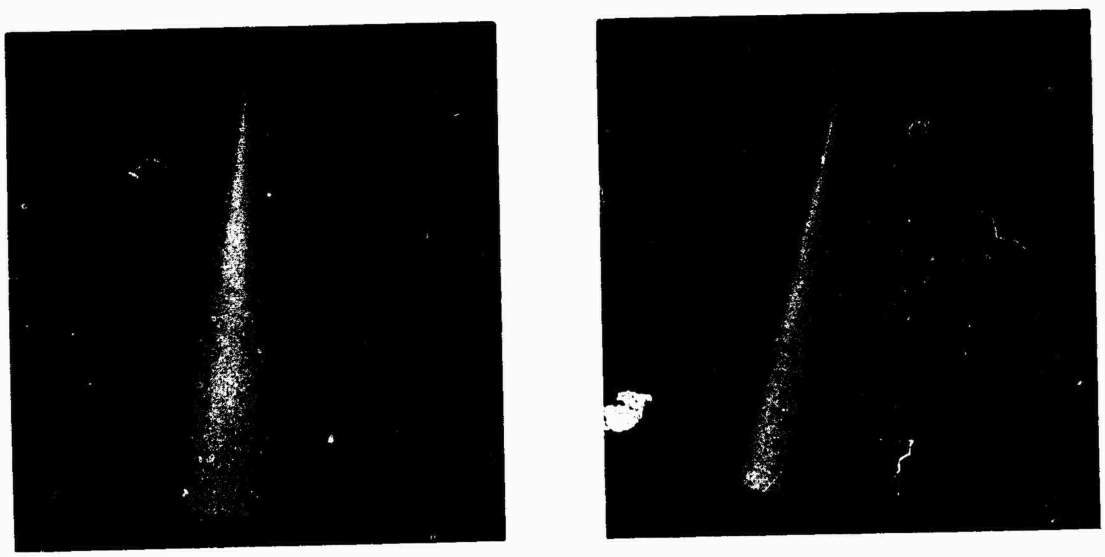


(b) flat white paint

Figure 4.15 Real cones with different light orientations and surface finishes

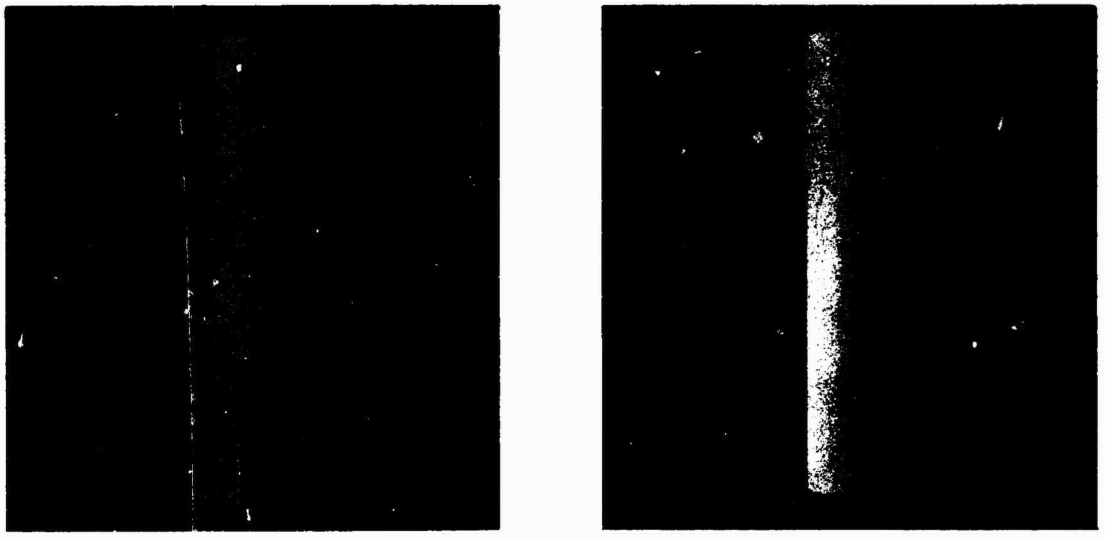


(a) high specular reflection

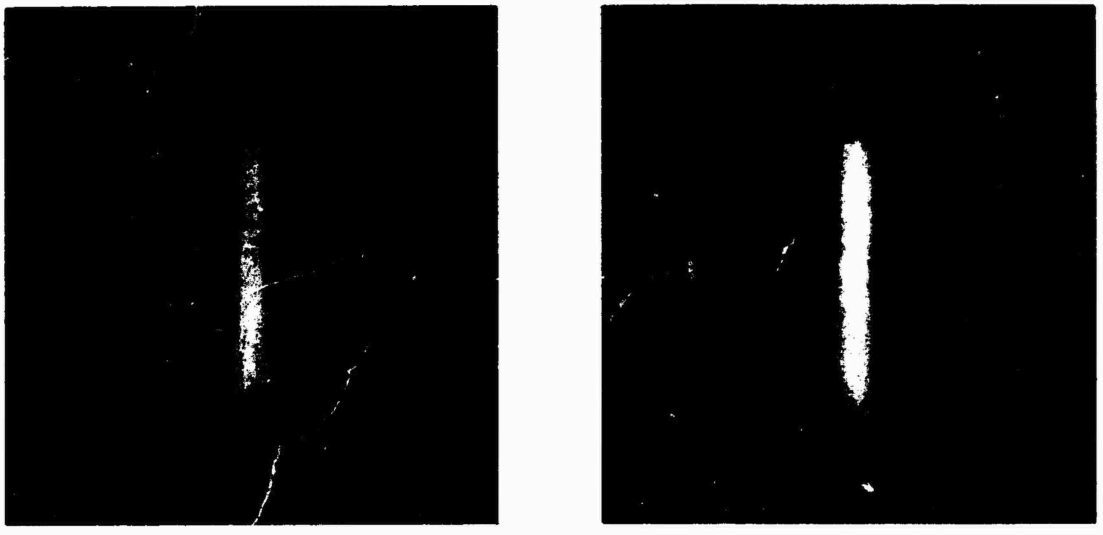


(b) low specular reflection

Figure 4.16 Cones rendered with improved shading technique

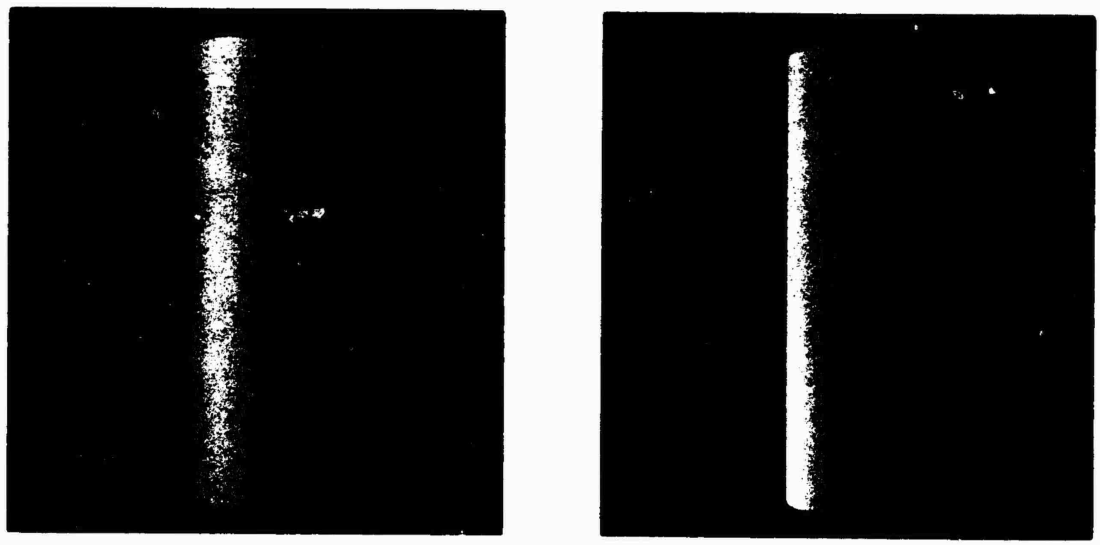


(a) flat white paint

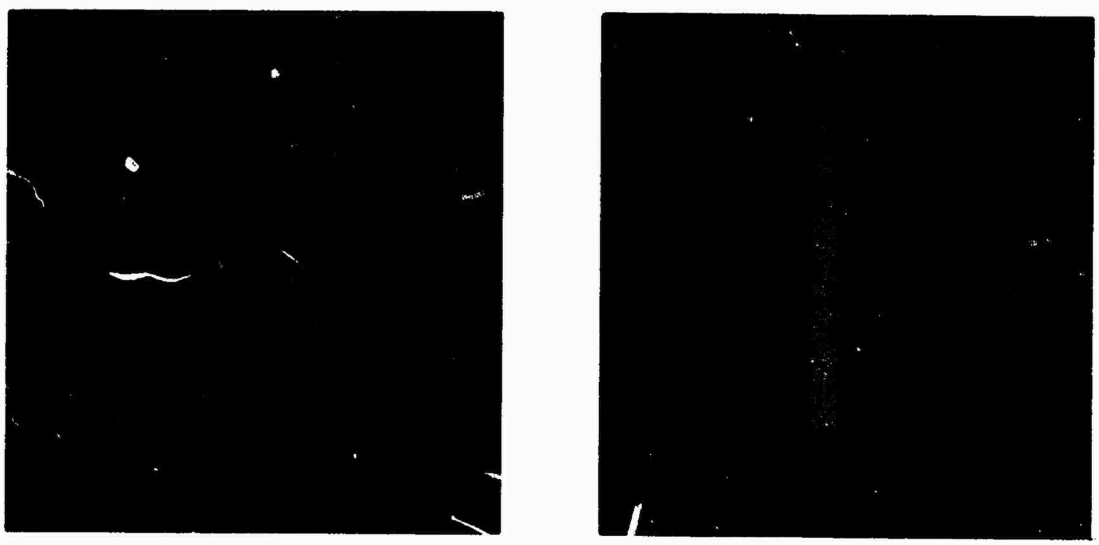


(b) aluminum paint

Figure 4.17 Real cylinders with
different light orientations and surface finishes



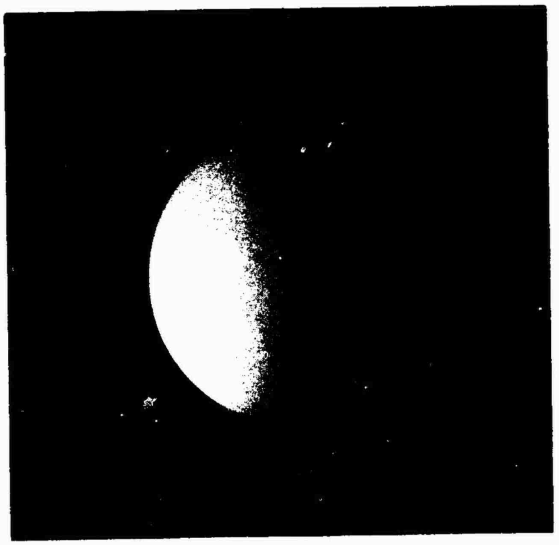
(a) low specular reflection



(b) high specular reflection

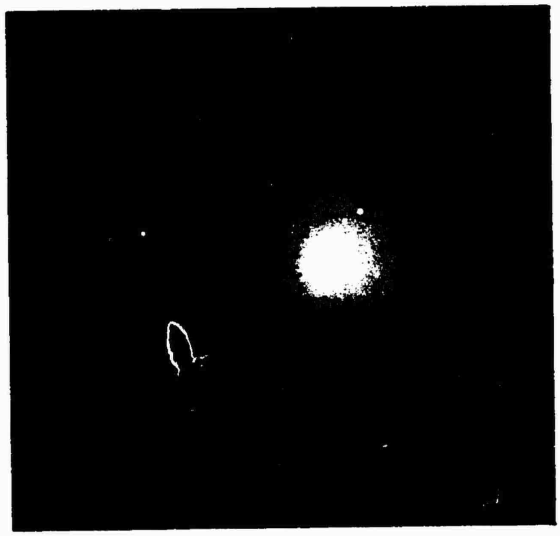
Figure 4.18 Cylinders rendered with improved shading technique

(a)

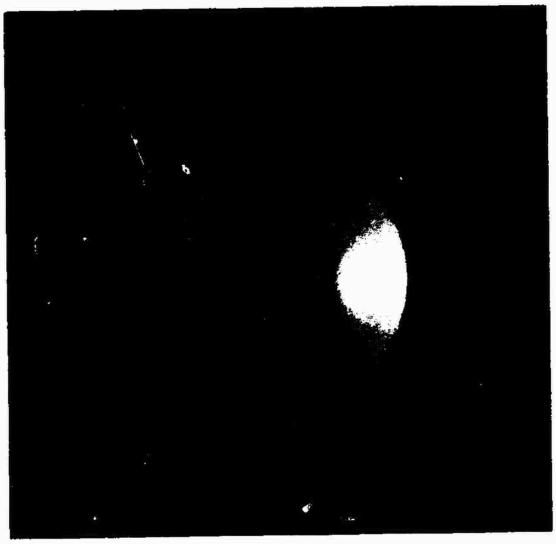


(b)

Figure 4.19 Real spheres with different light orientations



(a)



(b)

Figure 4.20 Spheres rendered with improved shading technique

W(i). Roughly, the ratio of the specular reflection of light simulating a highly reflected material like aluminum is about sixty percent of the incident white light. The power n in that case is four. For a flat white paint, the specular reflection coefficient is around ten percent, and the power n is equal to two.

Appendix II will give some values for the ratio W(i) and the power n for a sample of various materials.

IV. 5 Other interpolation schemes.

As was pointed out previously, the linear interpolation of the normal does not give a shading function with a continuous first derivative, thus a non-linear interpolation scheme may be desirable. Unfortunately, higher degree interpolation curves present some undesirable problems of additional conditions resulting in increased computation time and unsolvable problems due to the polygonal description model. Two possible non-linear interpolation schemes which exhibit these problems are presented here: (1) a spline under tension and (2) a cubic function.

The notion of spline under tension was introduced by Schweikert [28]. It is an attempt to imitate cubic splines, with the advantage of avoiding the production of extraneous inflection points in the curve. A spline under tension can be physically compared to a light and flexible bar which passes through the given points and which responds to the tension produced by pulling on its ends. The occurrence of extraneous inflection points can be then avoided by varying the tension.

A representation of the spline under tension is given by Cline [29],

which can be briefly described as follows:

Given a set of knots (xi) (with i=1,...,n such that xi<xi+1), a corresponding set of function values (yi) (with i=1,...,n) and a nonzero constant sigma, called the tension factor, then a function f which is defined as a spline under tension will satisfy the following conditions:

$$f(xi) = yi \qquad i=1,...,n \qquad (4.10)$$

and the quantity (f" - (sigma)↑2*f) varies linearly on each of the intervals [xi, xi+1] with i=1,...,n-1.

A mathematical solution to the previous equation is given by Cline. It requires the solving of a linear differential equation system.

In the case where sigma=0, the curve obtained is the usual cubic spline, and when sigma is very large, the solution to the system is a piecewise linear function. Several inconveniences arising due to the use of a spline under tension to interpolate the normal between adjacent vertices of a polygonal model in a real-time environment can be summarized as follows:

1. The linear differential equation system must be solved for each scan line. The present techniques used to solve a linear differential equation system are not fast enough for a real-time display system.

2. For a shading computation in the polygonal model, the normal to a point on an edge is interpolated from the vectors at the adjacent end points of the edge. This requires the interpolation of each individual component of the normal. In order to solve the differential equation system which defines a spline under tension, a second derivative of each component is required. An approximation to

the first derivative of a component can be obtained by using some techniques described by Bezier [30], but a second derivative for each component of the normal cannot be defined from existing information.

3. In the case where a spline under tension can be defined, it cannot be guaranteed that the curve will be free of extraneous inflection points, causing bumps and loops. These bumps and loops caused by the interpolant curve must be avoided in the computation of the shading.

4. The interpolation is done for each component of the normal. As the length of the normal at the vertices is of unit length, it is desirable that the length of the interpolated normal vector also is of unit length. However, since this is not obtained with any interpolation scheme, then a renormalization of the normal would be necessary.

5. A spline under tension cannot fit a set of given knots segment by segment. Every time an additional point is added, a new computation of the entire spline curve is required. This nonlocal property of the spline under tension causes many implementation problems.

The same difficulties encountered with a spline under tension arise when a cubic function is used to interpolate the normal for shading.

A cubic polynomial is defined as:

$$y = a*x\hat{}3 + b*x\hat{}2 + c*x + d \qquad (4.12)$$

The coefficients a, b, c, and d, are the solutions of a system of four linear equations. In order to solve this system of equations, four values of x and y are necessary.

Bezier pointed out two drawbacks of this method of representation:

1. The defined curve cannot have a vertical tangent.

2. The shape of the curve depends on the orientation of the coordinate axes.

In order to avoid these two inconveniences, a parametric representation of the curve is preferred. With the parametric representation, a point on an arc of a curve is represented by the expression:

$$x(t) = a*t\hat{}3 + b*t\hat{}2 + c*t + d$$

in which t is a parameter.

To define the four vectors a, b, c, and d, the following conditions must be satisfied:

1. t is equal to zero at the initial point.

2. t is equal to 1 at the final point.

3. The values of the coordinates at the initial and final points, as well as those of the parametric derivatives of the curve at the end points need to be given.

In this representation, the parametric derivative is represented by a vector which is tangent to the curve, but not simply by the direction of the tangent. The information for the computation of the length of the tangent vector is not available from the polygonal model.

In the case of a mathematically defined model, additional conditions can be supplied for a non-linear interpolation scheme. However, even with this

information, complex problems need to be solved.

CHAPTER V

HARDWARE IMPLEMENTATION

The described illumination model is independent of the hidden-surface algorithm. Since the determination of the shading uses a linear interpolation of the normal vector, only information on the components of the normal need to be handled by the hidden-surface in addition to that concerning the color and reflectance properties of the surfaces approximating the object.

The specularly reflected light is obtained from the normal at each point of the surface by an appropriate transformation which requires very little computation as described in the previous section. The transformation elements are computed once for each frame, and stored in the shader for the whole picture. The reflection curve W(i) for a reference material can be stored inside the shader itself. To obtain a specularly reflected light intensity corresponding to an angle of incidence i, and to a material the user desires to simulate, an access to the table of values for the reference reflection curve W(i) will give a value from which a proper intensity c be computed.

The shading algorithm is executed at the last step of the picture generation process. It works on the screen space of the cathode ray tube, and it is concerned only with the visible parts of the object.

In a polygonal description model, a normal vector is retained for each

vertex. When a output is done from scan line to scan line, the hidden-surface algorithm must update each component of the normal along edges which cross the scan line. In the display of a visible segment of a polygon on a scan line, the components of the normals at the ends of the segment are supplied to the shader. At that time, a linear interpolation of the normal for each resolution points which exist within the segment is done. Once the interpolated normal is obtained, it needs to be renormalized, so that it will have a unit length. As the coordinates of the normal are relative to the light source coordinate system, the component in the direction of light will be used in the computation for the diffuse reflected light intensity. A transformation of the unit length normal vector will determine the specularly reflected light intensity.

The most time consuming part of the algorithm is due to the normalization of the normal for each resolution point. This requires taking the square root of a number. A special purpose machine can be built for this purpose, and a sketch of a possible hardware implementation of part of the described algorithm is represented by the schematic diagram in Figure 5.1.

As shown by this diagram, all the operations requiring a multiplication or division of two numbers are performed by a lookup into read-only memory tables. When several operations can be done at the same time, they are executed in parallel. Also, the machine is synchronous, so that, except for the first point of each frame, the delay for computing the shading of each resolution point is one cycle of the machine. A cycle is determined by the execution time of the slowest element of the machine, which is the time needed to have access to the read-only memory. For each beginning point of a frame, a delay of about 15 machine cycles is needed as shown on the diagram. Since the access memory cycle is about 50

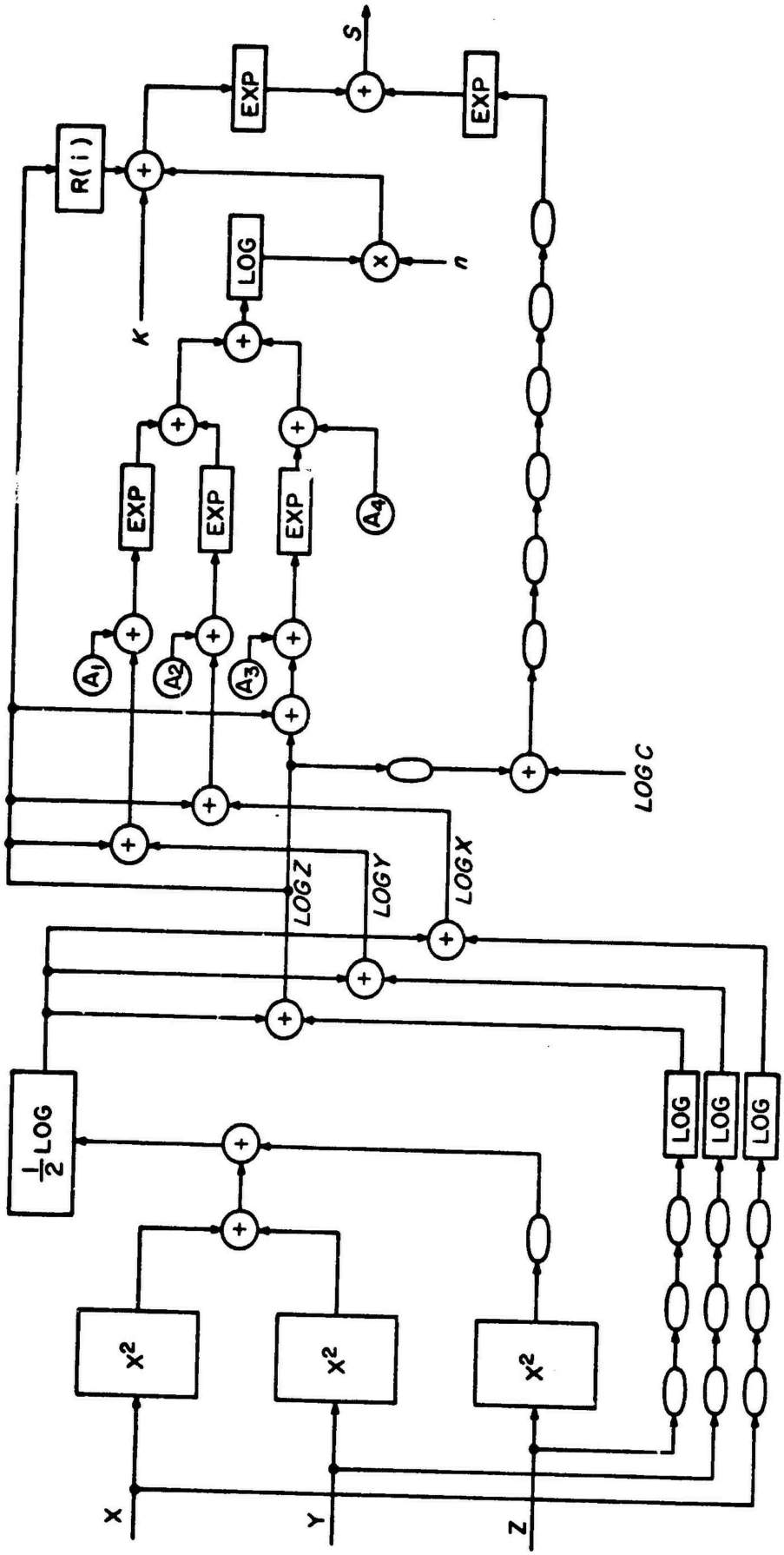


Figure 5.1 Schematic diagram of a possible hardware implementation

nanoseconds, a delay of about 1 microsecond is required. This is negligible when compared with the frame to frame delay.

The choice of lookup tables has been proposed because this solution is very fast and the price on the parts involved in the implementation are not expensive. Also, the logical design for this implementation is relatively simple.

This piece of hardware can be implemented alone or as the final output from a hidden-surface algorithm which can handle the information on the normals. As an example, the described illumination can be implemented with the Watkins Visible Surface algorithm for a real-time display system.

Watkins' algorithm stores the polygons as a set of pairs of adjacent edges. By sorting in the Y, X, and then Z-direction of the coordinate system, Watkins can find the visible polygon bounded by a pair of edges on a segment of the scan line. The visible segment is then sent over to the shader to be painted.

With the additional information about the normal at the starting point of each edge and the incrementation of its components along the edge, Watkins' algorithm can update this information from scan line to scan line. Therefore, an implementation of the described shading algorithm with the Watkins Visible Surface algorithm can be done without problem in hardware.

CHAPTER VI

CONCLUSION

Although the described illumination model can increase greatly the realism of computer-generated images, several important problems need to be solved before synthetic images can be really comparable to the real ones.

Some of these problems are due to the polygonal description model. Particularly, the silhouette of an object does not look smooth when the polygonal model is used to approximate a curved surface. This is also true for the implied edges, which result from the intersections of two surfaces. The linear approximation to a curved silhouette of the simulated object detracts from the smooth appearance of the object. At the present time, when a smooth profile of the object is desired, as in the case of the pictures made for comparison with pictures of real objects, a large numerical data base is required. Also the polygonal model does not allow the cutting of a polygon into different areas of distinct colors. This is desired when a simulation of a runway or a highway is performed: the stripes and numbers painted on the surface of the road belong to the road itself. If they are individually defined as sets of distinct polygons, the numerical data and the computation time for finding the visible surface will increase greatly.

The digital sampling technique introduces two more problems: 1) Due to the raster points, a straight line close to a vertical or horizontal line is very

difficult to draw. Usually, the resulting line is not straight, but it looks like it is made of a succession of parallel segments. 2) When a long and thin polygon is displayed, as in the case of the lane separation on a runway, the polygon can disappear if its size in one direction is smaller than the resolution of the screen. This effect is annoying in the case of a movie, where a polygon is present in one frame but has vanished in the next frame. The worse case arises when a line is broken into disjointed segments. One of the techniques which has been used to decrease the first effect is the mixing of the color of the adjacent polygons along their common boundary.

Finally a greater improvement of the quality of the synthetic pictures will be obtained when shadows and the texture of the objects can be simulated. When solutions to these problems are found, a new world will be opened in the domain of computer-generated images.

BIBLIOGRAPHY

1. Roberts, L. G. "Machine Perception of Three-dimentional Solids," M.I.T. Lincoln Laboratory, Cambridge, Mass., TR 315, May 22, 1963. Also in Optical and Electro-optical Information Processing, Tipper et al, eds. M.I.T. Press, p. 159.

2. Sutherland, I. E., Sproull, R. F., and Schumacker, R. A. "A Characterization of Ten Hidden-Surface Algorithms," Proc. of the National Computer Conference, New York, June 1973.

3. MAGI, Mathematical Applications Group Inc. "3-D Simulated Graphics," Datamation, 14, February 1968, p. 69.

4. Comba, P. G. "A Procedure of Detecting Intersections of Three-Dimentional Objects," Report 39,020, I.B.M. New York Scientific Center, January 1967.

5. Weiss, R. A. "Be Vision, a package of I.B.M. 7090 Fortran Programs to Draw Orthographic Views of Combinations of Planes and Quadric Surfaces," JACM, 13, April 1966, pp. 194-204.

6. Mahl, R. "Visible Surface Algorithm for Quadric Patches," IEEE, TC-21, P.1, January 1972.

7. Catmull, E. E. "An Algorithm for Visible Surface Display." Ph.D. thesis, Department of Computer Science, University of Utah (To appear).

8. Clark, J. H. "Computer-Aided Design of Free-Form Surfaces," Ph.D. thesis, Department of Computer Science, University of Utah (To appear).

9. Galimberti, R., and Montanari, U. "An Algorithm for Hidden Line Elimination," CACM 12, 4, April 1969, p. 206.

10. Kubert, B. R. "A Computer Method for Perspective Representation of Curves and Surfaces," Aerospace Corporation, San Bernadino Operations, April 1969.

11. Loutrel, P. P. "A Solution to the Hidden-Line Problem for Computer-Drawn Polyhedra," NYU Engineering and Science, Department of Electrical Engineering Report 400-167, September 1967. Also in IEEE Transactions on Computers EC-19[3], March 1970.

12. Romney, G. W. "Computer Assisted Assembly and Rendering of Solids," Department of Computer Science, University of Utah, TR 4-20, 1967.

13. Warnock, J. E. "A Hidden-Line Algorithm for Halftone Picture Representation", Department of Computer Science, University of Utah, TR 4-15, 1969.

14. Watkins, G. S. "A Real-Time Visible Surface Algorithm," Department of Computer Science, University of Utah, UTECH-CSc-70-101, June 1970.

15. Bouknight, W. J. "A Procedure for Generation of Three-dimentional Half-Toned Computer Graphics Representations," CACM 13, 9, September 1969, p. 527.

16. Kelley, K. C. "A Computer Graphics Program for the Generation of Half-Tone Images with Shadows," Universtiy of Illinois, Coordinated Science Lab., R-444, November 1969.

17. Appel, A. "Shadows Without Substance: Some Techniques for Shading Machine Renderings of Solids," IBM Research Center, Yorktown Heights, New York.

18. Rougelot, R. S., and Schumacker, R. A. "General Electric Real-Time Display," NSA Report, NAS 9-3916.

19. Gouraud, H. "Computer Display of Curved Surfaces," Department of Computer Science, University of Utah, UTEC-CSc-71-113, June 1971. Also in IEEE, TC-20, June 1971, p. 623.

20. Newell, M. E., Newell, R. G., and Sancha, T. L. "A New Approach to the Shaded Picture Problem," Proc. of the ACM 1973 National Conference.

21. Cornsweet, T. N. "Visual Perception." Academic Press, New York, 1970.

22. Bekesy, G. von. "Neural Inhibitory Units of the Eye and Skin. Quantitative Description of Contrast Phenomena." S. Opt. Soc. Am., 50, p.p. 1060-1070.

23. Hartline, H. K. "Inhibition of Activity of Visual Receptors by Illuminating Near By Retinal Elements in the Limulus Eye." Fed. Proc., 8, 69, 1949.

24. Ratliff, F. "MACH BANDS: Quantitative Studies on Neural Networks in the Retina." Holden-Day Inc., San Francisco, 1965.

25. Stockham, T. G. "Image Processing with Context of a Visual Model," in Proc. of the IEEE., 60, 7, July 1972, pp. 828-842.

26. Jenkins, F. A., and White, H. E. "Fundamental of Optics." McGraw-Hill Book Company Inc., New York, 1957.

27. Evans, R. M. "An Introduction to Color." John Wiley & Sons Editor, New York, 1948.

28. Schweikert, D. G. "An Interpolation Curve Using a Spline Under Tension." Journal of Mathematics and Physics, 45, 1966, pp. 312-317.

29. Cline, A. K. "Curve Fitting in One- and Two-Dimentional Spaces Using Splines Under Tension." National Center for Atmospheric Research, Boulder, Colorado.

30. Bezier, P. "NUMERICAL CONTROL: Mathematics and Applications." John Wiley & Sons, New York, 1972.

31. Stockham, T. G., "Photographic Processing: Photographic Display Techniques." Unpublished Document, Department of Computer Science, University of Utah.

32. Luckiesh, M. "Light and Shade." D. van Nostrand Company, New York, 1916.

APPENDIX I

COMPENSATION TABLE

The production of a synthetic picture is a long process which involves the numerical description of the object model, the processing of this numerical data and the display of the final image on a cathode ray tube. The final display step is not the least important, because the subjective quality of the generated pictures will depend on it.

The diagram in Figure A1.1 shows the different steps along the display process to record a generated-image on film. It can be briefly described as following:

The display program takes as input a numeric data description of a scene, and will compute a brightness n for each part of the scene corresponding to a resolution unit of the scope. This number is then corrected by a table of values C. The corrected output z is then processed by the digital to analog converter (D.A.C.). The output voltage Z from the D.A.C. is used to excite the z-axis amplifier of the scope. The two values b and c allow adjustment, respectively, of the initial brightness and the contrast of the cathode ray tube. These two values change only when the characteristics of the scope or the film change, otherwise they remain constant during the display time. The phosphor of the screen radiates a certain light energy L when a voltage tension n is applied to the scope amplifier, the film will be then impressed with a certain light

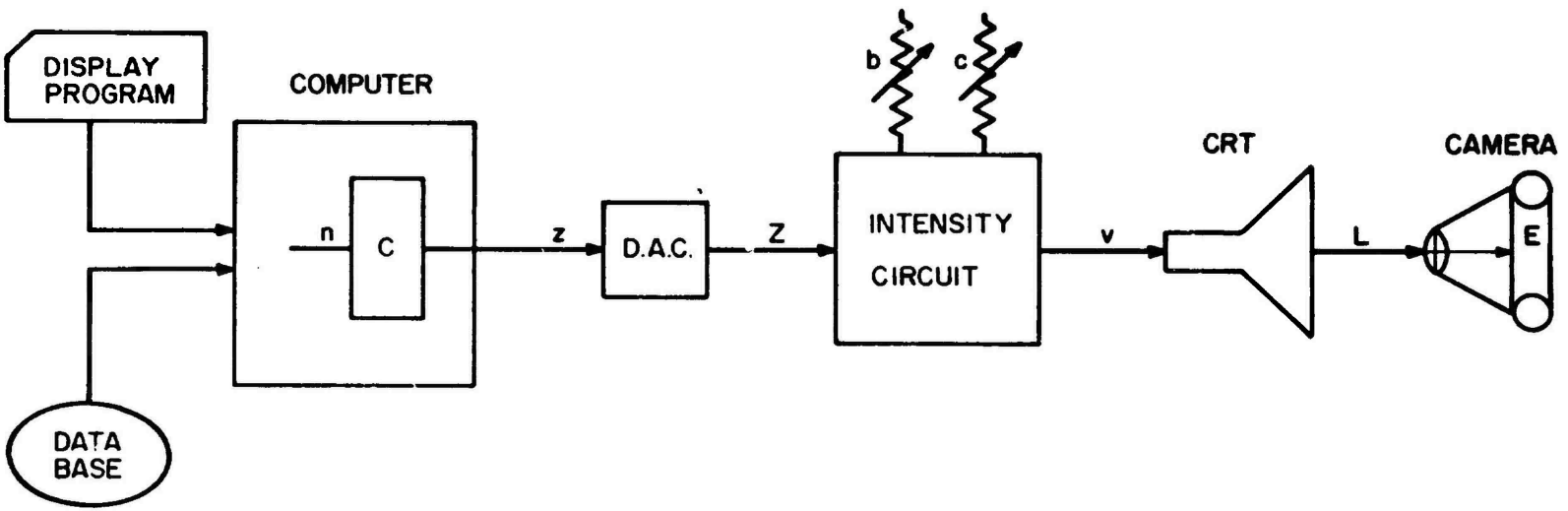


Figure A1.1 Computer display system

energy E.

The transformation from a numerical value to a subjective brightness is a complex task, because it involves several systems: the electronics of the scope, the phosphor of the cathode ray tube and the subjective brightness perception of the eye. As none of these systems are linear, it is then necessary to define a function which can transform numerical data into brightness levels.

Among the non-linearities of the systems involved in the generation of a synthetic image, only the non-linearity of the photographic film cannot be mathematically formulated; therefore a technique has been developed to compensate for this non-linearity (Stockham [31]).

The cathode ray tube phosphor can be chosen so that the light intensity radiated by the screen is a linear function of the energy of the electron beam striking the screen. This energy, therefore the light intensity, can be approximated by the following function:

$$I = d*(vc-vo)\uparrow Gamma$$

where d is a constant, vo is the excitation voltage corresponding to an absence of light on the screen, and vc is the actual excitation voltage.

Knowing the power Gamma, it is then possible to produce a linear relation between n and the radiated light energy L by using a compensation table for this value of Gamma. In most of the cases, Gamma is about .33.

Assuming now that the phosphor of the screen of the CRT is linear and that the non-linearity of the z-axix amplifier of the scope can be compensated with a Gamma table, it is then desired that the reflectance (or transmittance in

the case of a transparency) of a photographic material be linearly proportional to the incident light energy received by this material. In practice this is not possible. However, it is known that the incident energy E is a linear function of the value n, and in photography, the density D is defined as:

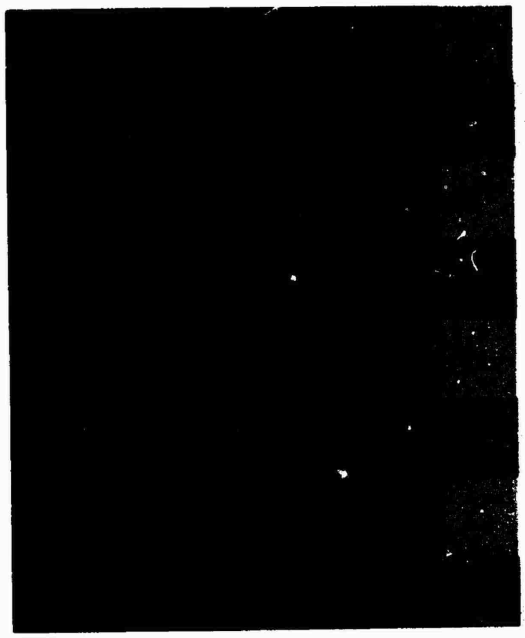
$$D = \log (Ir)$$

where Ir is the ratio between the incident light I and the relected light R.

It is then possible to have a representation of the sensibility function of the photographic material by plotting the density image against the logarithm of the incident energy. Such a curve can be obtained more or less empirically in the following way:

The initial brightness b of the scope is adjusted so that a zero value of z corresponds to the cut-off of the image intensity on the screen. This is done in a completely dark room after letting the eye acclimate to the darkness for several minutes. When the initial brightness is well adjusted, a greyscale like the one called stepwedge in Figure A1.2a is displayed with a Gamma correction. This greyscale is then recorded on a photographic material which requires compensation. The values of c, the f-stops of the camera lens and the exposure time are adjusted so that a maximal use of the capabilities of the D.A.C., the phosphor of the screen and the dynamic range of the film can be obtained.

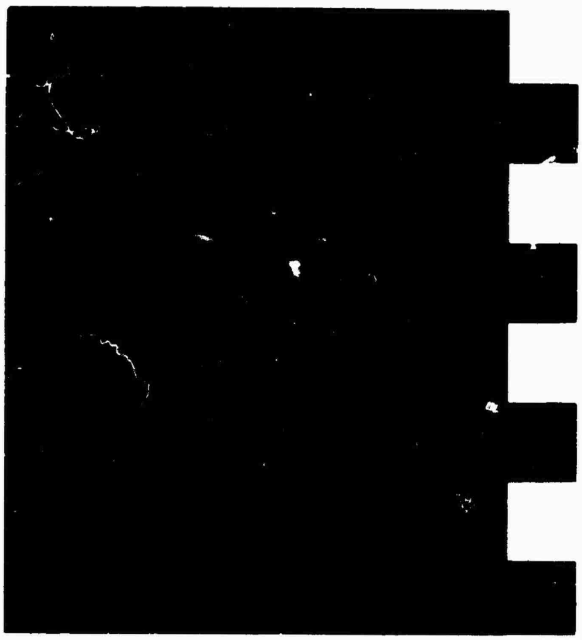
From a similar stepwedge dispayed with a Gamma compensation (Figure A1.2a), the densities for each square on the stepwedge are measured with a densitometer and their values plotted against the corresponding values of log (n). A curve is then fitted to this set of points by using either a spline fitting scheme or a polynomial function fitting. A such density curve for the black and



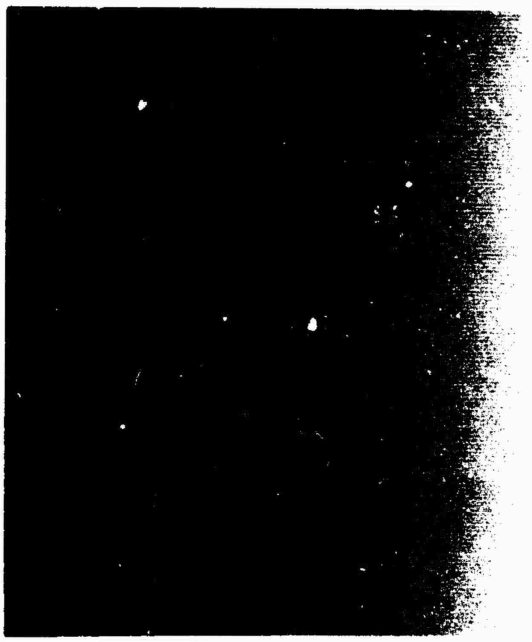
(a) Stepwedge

(b) Linear intensity plot

Figure A1.2 Linear density



(a) stepwedge



(b) linear intensity plot

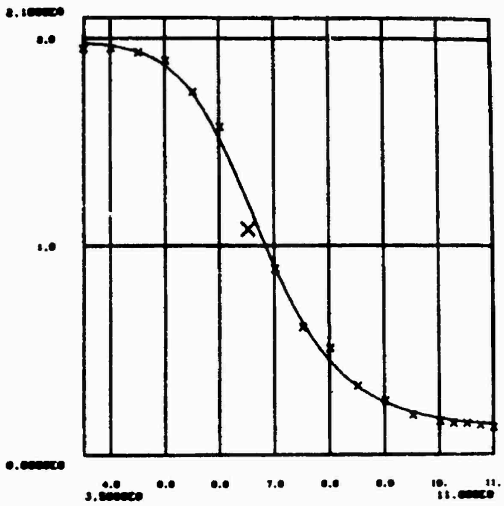
Figure A1.3 Linear subjective brightness with specular reflection

white Polaroid type 52 is shown in Figure A1.4a. Knowing the value of Gamma, a final compensation curve can be obtained from the film correction (Figure A1.4b).

The new stepwedge using the final compensation table is then displayed in Figure A1.2a. The value of n corresponding to each square is divided each time by 2 when going right to left from one square to the next in a row. The corresponding densities will be linearly increasing.

This compensation for the non-linearity of a photographic material assumes that for an input intensity n, the output intensity R is a linear function of n. However in the production of a synthetic image, the shading of each point of the object model is computed as a brightness level. This brightness is called a subjective brightness and has been defined in Chapter II on "Visual Perception" as a linear function of the logarithm of the intensity. Therefore, when a shading n1 is computed for a white material, it is desired that the simulated material look white in the generated image. For a black material, a shading value n2 is obtained and it is desired that the image of this material be as black as possible. For a shading n3 = (n1/2) corresponding to a grey color, half way between black and white, it is desired that the corresponding image be neutral grey as expected.

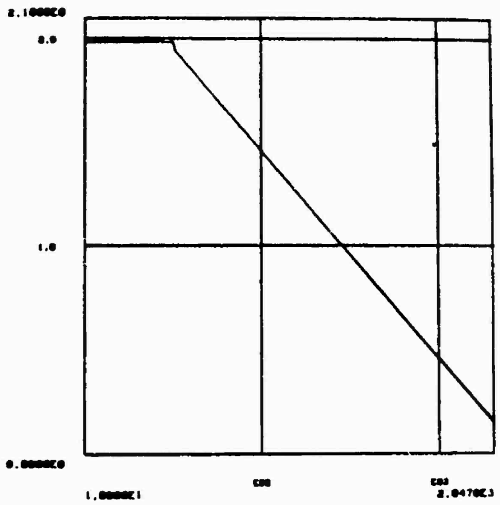
With the previous compensation table, the entry log (n1/2) will give a density value which is not the mean density for black and white densities, because the table gives the density as a linear function of the logarithm of n. As the subjective brightness is a linear function of the density, the subjective brightness corresponding to n3 = (n1-n2) / 2 is not grey but actually appears much lighter.



(a)

Density curve for black and white

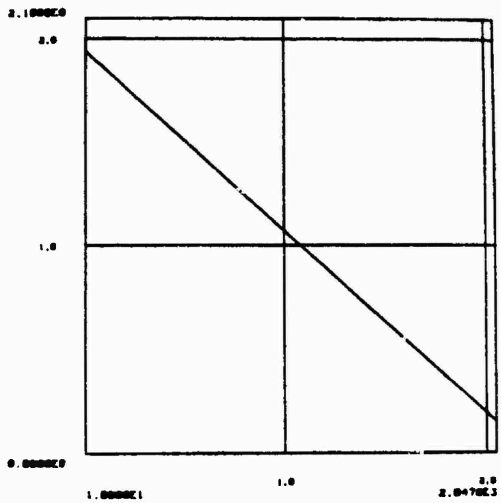
Polaroid Type 52



(b)

Compensated curve for

B & W Polaroid with Gamma=.34



(c)

Subjective brightness

compensated curve

Figure A1.4 Compensation curves

A linear plot of the value n against the density would give a linear increase in the light intensity from a darkest point to a lightest point. This linear intensity plot using the described compensation table will give the brightness effect in Figure A1.2b. In this illustration, most of the area appears to be very light, and toward a small value of n, the intensity decreases quickly to its lowest level in a short range of the value n.

In order to have a linear increase of subjective brightness as a function of n (Figure A1.3b), a table to compensate for the non-linearity of the film must be obtained having the density image vary linearly with the value of n. A stepwedge displayed in Figure A1.3a is made with this table. The two compensation curves for these tables are displayed in Figure A1.4b and A1.4c.

The generated pictures illustrated in this thesis incorporate a compensation table which gives a linear function between the density image and the value n.

The compensation table provides only for black and white film, because color phtographic materials involve deeper research in crosstalks between different color primaries of the film. Intensive research in this area is being done at the University of Utah, under the guidance of Professor Thomas Stockham, Jr.

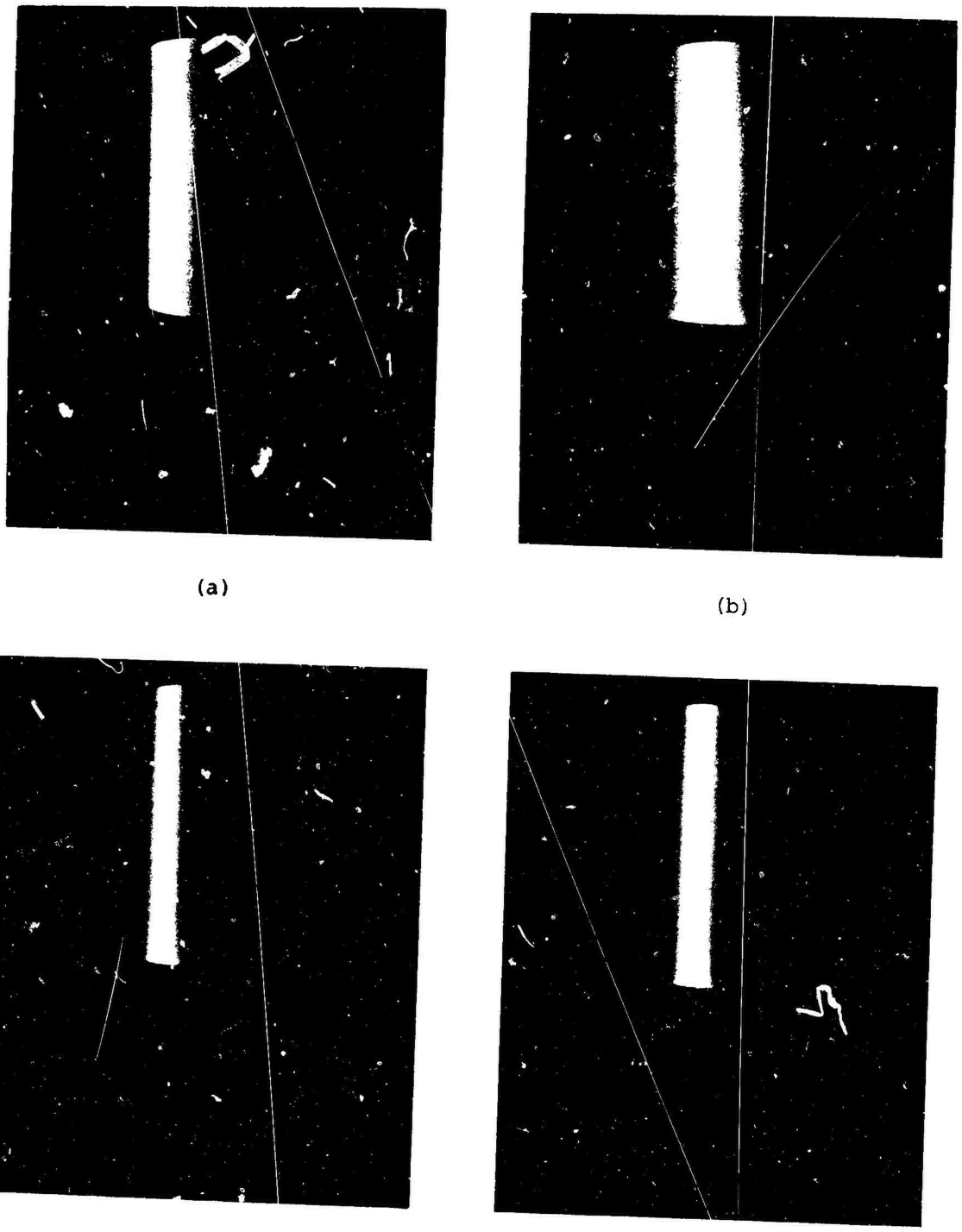
APPENDIX II

VALUES OF SOME REFLECTION COEFFICIENTS

The author attempted to determine empirically the reflection curves as defined in chapter III ("Some Physical Aspects of Light and Shade") for a set of sample materials. The experiment was set up in an environment excluding diffuse and indirectly reflected light. A point light source approximated by a ziconium arc provided the only incident light. A material to be analyzed is illuminated by the light source at different incident angles. The reflected light intensity was then measured with a very sensitive lightmeter. Also the values of the reflected light close to the reflection angle are recorded so that not only the reflection curve W(i) could be determined but the power n (see equation 4.5) could also be fixed.

Unfortunately, this experiment has not been successful due to the difficulty of obtaining sufficient incident light intensity. The equipment was further hampered by a lack a sensitivity in the light measuring device.

A similar experiment has been done by Professor Ronald Resch at the Universty of Utah, but the incident light used in that case was an ordinary light bulb and the diffuse and reflected light due to the environment were not completely eliminated. The results of this experiment showed that W(i) is not constant, but they were not precise enough to allow a determination of the shape of the curve W(i).



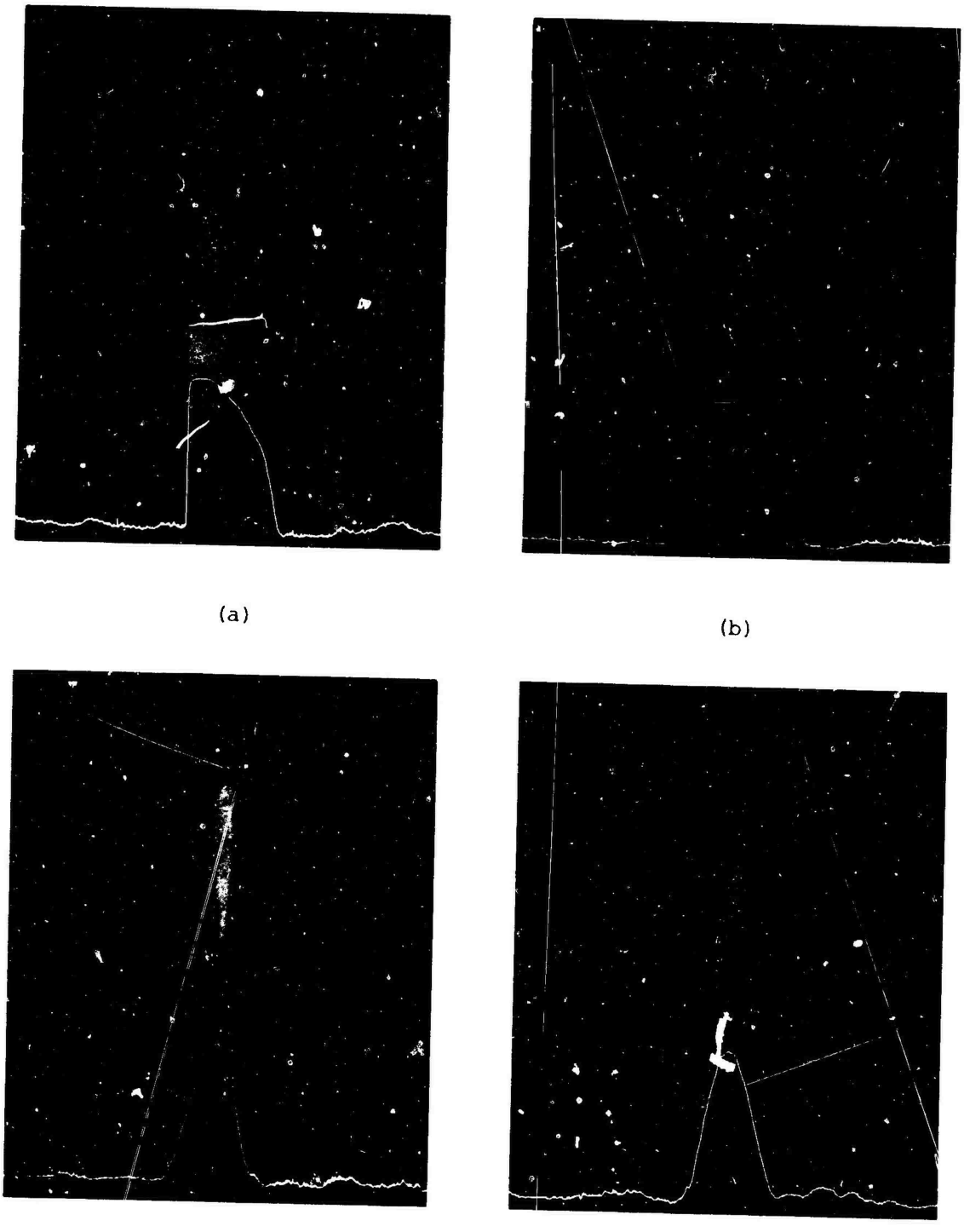
(a)

(b)

(c)

(d)

Figure A2.1 Rendered cylinders and related light intensity curves



(a) (b)

(c) (d)

Figure A2.2 Real cylinders and related light intensity curves

An attempt was made to use an optical scanner to digitize pictures of real objects painted with diffrent surface finishes; and, from the data obtained, to determine the reflection curves for the simulated material. This technique is too time consuming (about 20 minutes of computer CPU time of the DEC PDP-10 for digitizing each picture) and complicated (compensation must be done for each individual picuture scanned in). Therefore, only one set of pictures have been digitized and their reflection curves generated (Figure A2.1) for comparison with synthetic pictures and their related reflection curves (Figure a2.2). As shown by these curves, the distribution of light intensity in an image of a real object can be identically reproduced with the improved shading technique. This is obtained by adjusting the reflection coefficient Cp and the ratio W(i) of the equation 4.5. However, there is no scientific technique which would allow a precise measurement of these values.

In the case of a perfectly diffuse material, an approximative determination of the reflection coefficient Cp may be obtained. Luckiesh [32] gave a set of values of Cp for diffuse materials of different colors. A part of it is reproduced here as an example:

White diffusing	.80
Red-purple	.16
Red	.2
Orange	.38
Yellow	.60
Yellow-green	.46
Saturated green	.32
Blue	.23

Violet .14

To simulate highlights, it is necessary then to add W(i) to these values. In order to generate the illustrations of this thesis, it was found that for a diffuse material, W(i) was about .2 and the power n was equal to 2. In the case of a highly reflected material like polished metal, W(i) ranged from .6 to .8 and n was raisen to 5 or 6.

ACKNOWLEDGEMENTS

I am indebted to Professor David Evans for his constant help and his enthusiastic support without which this project could not have been completed. I would like to thank Professor Ivan Sutherland for his helpful advice in the development of this project. I would like also to thank Professor Richard Riesenfeld for aiding in the correction of my somewhat balky English.

Many thanks to my fellow students, particularly James Clark with whom I spent hours of fruitful discussion, Franklin Crow who is responsible for the beautifull pictures of tranparent objects, and to George Randall who shared many desperate hours building a "correct" correction table. Thanks also to the system group, particularly to Dennis Ting who tried hard to keep the machines up and to John Riley who helped in the correction of the draft of this thesis.

A special thanks to Mike Milochik for his great artwork in all of the accompanying illustrations.