CHAPTER

The three-dimensional world

16

Humans are able to employ 3-D vision with consummate ease, and according to conventional wisdom, binocular vision is the key to this success. The truth is more complex than this, and this chapter demonstrates why.

Look out for:

- What can be achieved using binocular vision
- How the shading of surfaces can be used in place of binocular vision to achieve similar ends
- How these basic methods provide dimensional information for 3-D scenes but do not immediately lead to object recognition
- How the process of 3-D object recognition can be tackled by studies of 3-D geometry.

Note that this is an introductory chapter on 3-D vision, designed to give the flavor of the subject and to show its origins in human vision. It will be followed by the other four chapters (Chapters 17–20) that comprise Part 4 of this volume.

At a more detailed level, notice the importance of the epipolar line approach in solving the correspondence problem: the concept is deservedly taken considerably further in Chapter 19, Image Transformations and Camera Calibration, in conjunction with the required mathematical formulation.

16.1 INTRODUCTION

In the foregoing chapters, it has generally been assumed that objects are essentially flat and are viewed from above in such a way that there are only three degrees of freedom—namely, the two associated with position, and a further one concerned with orientation. Although this approach was adequate for carrying out many useful visual tasks, it is inadequate for interpreting outdoor or factory scenes or even for helping with quite simple robot assembly and inspection tasks. Indeed, over the past few decades, a considerable amount of quite sophisticated theory has been developed and backed up by experiment to find how scenes composed of real 3-D objects can be understood in detail.

Computer Vision. DOI: http://dx.doi.org/10.1016/B978-0-12-809284-2.00016-2

In general, this means attempting to interpret scenes in which objects may appear in totally arbitrary positions and orientations—corresponding to six degrees of freedom. Interpreting such scenes, and deducing the translation and orientation parameters of arbitrary sets of objects, takes a substantial amount of computation—partly because of the inherent ambiguity in inferring 3-D information from 2-D images.

A variety of approaches are now available for proceeding with 3-D vision. A single chapter will be unable to describe all of them but the intention here is to provide an overview, outlining the basic principles and classifying the methods according to generality, applicability, etc. Although computer vision need not necessarily mimic the capabilities of the human eye−brain system, much research on 3-D vision has been aimed at biological modeling. This type of research shows that the human visual system makes use of a number of different methods simultaneously, taking appropriate cues from the input data and forming hypotheses about the content of a scene, progressively enhancing these hypotheses until a useful working model of what is present is produced. Thus, individual methods are not expected to work in isolation: rather, they need to provide the model generator with whatever data become available. Clearly, biological machinery of various types will lie idle for much of the time until triggered by specific input stimuli. Computer vision systems are currently less sophisticated than this and tend to be built on specific processing models, so that they can be applied efficiently to more restricted types of image data. In this chapter, we adopt the pragmatic view that particular methods need to be (or have been) developed for specific types of situation, and that they should be used only when appropriate—although some care is taken to elucidate what the appropriate types of application are.

16.2 THREE-DIMENSIONAL VISION—THE VARIETY OF METHODS

One of the most obvious characteristics of the human visual system is that it employs two eyes, and it is well known to the layman that binocular (or "stereo") vision permits depth to be discerned within a scene. However, the loss of vision that results when one eye is shut is relatively insignificant and is by no means a disqualification from driving a car or even an airplane. On the contrary, depth can readily be deduced in monocular vision from a plethora of cues that are buried in an image. Naturally, to achieve this the eye−brain system is able to call on a huge amount of prestored data about the physical world and about the types of object in it, be they manmade or natural entities. For example, the size of any car being viewed is strongly constrained; likewise, most objects have highly restricted sizes, both absolutely and in their depths relative to their frontal dimensions. Nevertheless, in a single view of a scene, it is normally impossible to deduce

absolute sizes—all the objects and their depths can be scaled up or down by arbitrary factors and this cannot be discerned from a monocular view.

Although it is clear that the eye−brain system makes use of a huge database relating to the physical world, there is much that can be learnt with negligible prior knowledge, even from a single monocular view. The main key to this is the "shape from shading" concept. For 3-D shape to be deducible from shading information (i.e., from the gray-scale intensities in an image), something has to be known about how the scene is lit—the simplest situation being when the scene is illuminated by a single point light source at a known position: note that indoors a single overhead tungsten light is still the most usual illuminator, whereas outdoors the sun performs a similar function. In either case, an obvious result is that a single source will illuminate one part of an object and not another—which then remains in shadow—and parts that are orientated in various ways relative to the source and the observer appear with different brightness values, so that orientation can in principle be deduced. In fact, as will be seen below, deduction of orientation and position is not at all trivial and may even be ambiguous. Nevertheless, successful methods have been developed for carrying out this task. One problem that often arises is that the position of the light source is unknown but this information can generally be extracted (at least by the eye) from the scene being examined, so a bootstrapping procedure is then able to unlock the image data gradually and proceed to an interpretation.

Although these methods enable the eye to interpret real scenes, it is difficult to say quite to what degree of precision they are carried out. With computer vision, the required precision levels are liable to be higher, although the machine will be aided by knowing exactly where the source of illumination is. However, with computer vision we can go further and arrange artificial lighting schemes that would not appear in nature, so the computer can acquire an advantage over the human visual system. In particular, a set of light sources can be applied in sequence to the scene—an approach known as photometric stereo—which can in certain cases help the computer to interpret the scene more rigorously and efficiently. In other cases, structured light may be applied: this means projecting onto the scene a pattern of spots or stripes, or even a grid of lines, and measuring their positions in the resulting image. By this means, depth information can be obtained much as for pairs of stereo images.

Finally, a number of methods have been developed for analyzing images on the basis of readily identifiable sets of features. These methods are the 3-D analogs of the graph matching and Generalized Hough Transform (GHT) approaches of Chapter 11, The Generalized Hough Transform. However, they are significantly more complex because they generally involve six degrees of freedom in place of the three assumed throughout Chapter 11, The Generalized Hough Transform. It should also be noted that such methods make strong assumptions about the particular objects to be located within the scene. In general situations, it is unlikely that such assumptions could be made, and so initial analysis of any images must be made on the basis that the entire scene must be mapped out in 3-D, then 3-D

models built up and finally deductions must be made by noticing what relation one part of the scene bears to another part. Note that if a scene is composed from an entirely new set of objects, all that can be done is to *describe* what is present and say perhaps what the set most closely *resembles*: recognition per se cannot be performed. Notice that scene analysis is—at least from a single monocular image—an inherently ambiguous process: every scene can have a number of possible interpretations and there is evidence that the eye looks for the simplest and most probable explanation rather than an absolute interpretation. Indeed, it is underlined by the many illusions to which the eye−brain system is subject that decisions must repeatedly be made concerning the most likely interpretation of a scene and that there is some risk that its internal model builder will lock on to an interpretation or part-interpretation that is suboptimal (see the paintings of Escher!).

This section has indicated that methods of 3-D vision can be categorized according to whether they start by mapping out the shapes of objects in 3-D space and then attempt to interpret the resulting shapes, or whether they try to identify objects directly from their features. In either case, a knowledge base is ultimately called for. It has also been seen that methods of mapping objects in real space include monocular and binocular methods, although structured lighting can help to offset the deficiencies of employing a single "eye." Laser scanning and ranging techniques must also be included in methods of 3-D mapping, although space precludes detailed discussion of these techniques in this book.

16.3 PROJECTION SCHEMES FOR THREE-DIMENSIONAL VISION

It is common in engineering drawings to provide three views of an object to be manufactured—the plan, the side view, and the elevation. Traditionally, these views are simple orthographic (nondistorting) projections of the object—that is, they are made by taking sets of parallel lines from points on the object to the flat plane on which it is being projected.

However, when objects are viewed by eye or from a camera, rays converge to the lens and so images formed in this way are subject not only to change of scale but also to perspective distortions (Fig. 16.1). This type of projection is called perspective projection, although it includes orthographic projection as the special case of viewing from a distant point. Unfortunately, perspective projections have the disadvantage that they tend to make objects appear more complex than they really are by destroying simple relationships between their features: thus, parallel edges no longer appear parallel and midpoints no longer appear as such (although many useful geometric properties still hold—e.g., a tangent line remains a tangent line and the order of points on a straight line remains unchanged).

In outdoor scenes, it is very common to see lines which are known to be parallel apparently converging toward a vanishing point on the horizon line (Fig. 16.2).

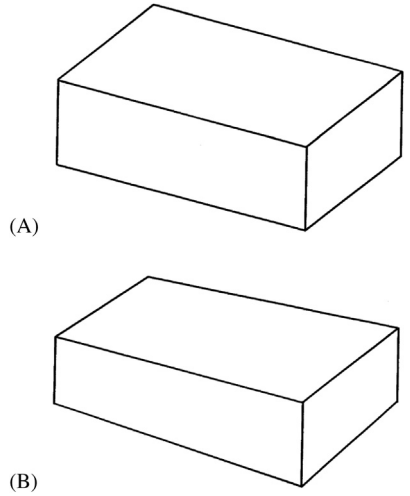

FIGURE 16.1

(A) Image of a rectangular box taken using orthographic projection; (B) the same box taken using perspective projection. In (B) note that parallel lines no longer appear parallel, although paradoxically the box appears more realistic.

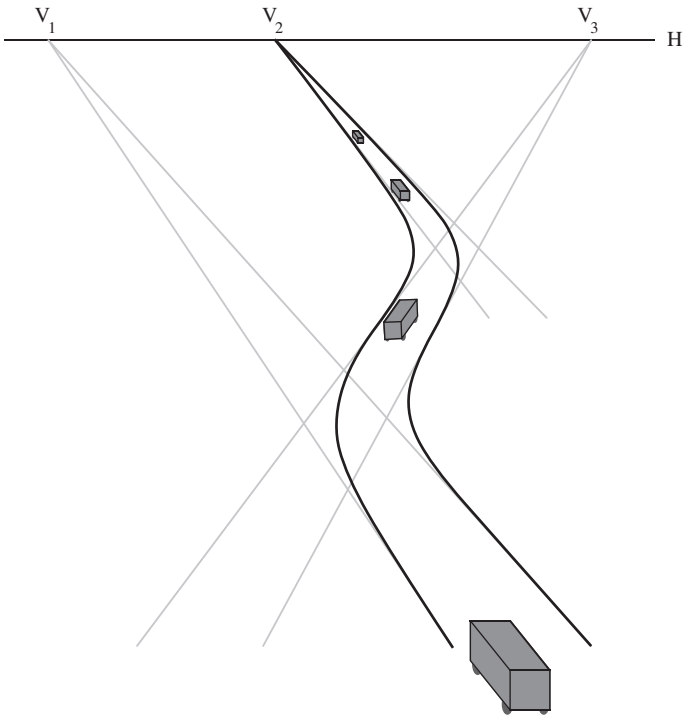

FIGURE 16.2

Vanishing points and the horizon line. This figure shows how parallel lines on the ground plane appear, under perspective projection, to meet at vanishing points V_i on the horizon line H. (Note that V_i and H lie in the *image* plane.) If two parallel lines do not lie on the ground plane, their vanishing point will lie on a different vanishing line. Hence, it should be possible to determine whether any roads are on an incline by computing all the vanishing points for the scene.

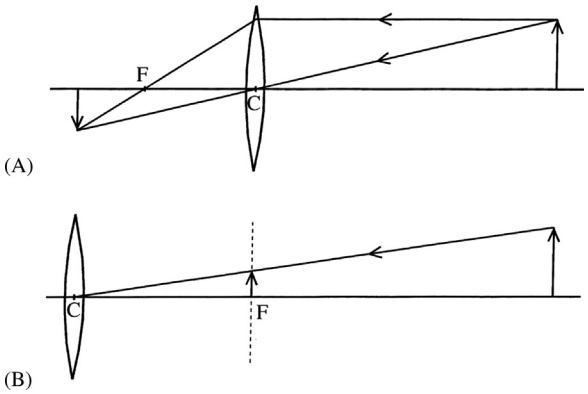

FIGURE 16.3

(A) Projection of an image into the image plane by a convex lens; note that a single image plane only brings objects at a single distance into focus but that for far-off objects the image plane may be taken to be the focal plane, a distance *f* from the lens; (B) a commonly used convention which imagines the projected image to appear noninverted at a focal plane *F* in front of the lens. The center of the lens is said to be the center of projection for image formation.

In fact, the horizon line is the projection onto the image plane of the line at infinity on the ground plane G: it is the set of all possible vanishing points for parallel lines on G. In general, the vanishing points of a plane P are defined as the projections onto the image plane corresponding to points at infinity in a given direction on P. Thus, any plane Q within the field of view may have vanishing points in the image plane, and these will lie on a vanishing line which is the analog of the horizon line for Q.

Fig. 16.3A shows how an image is projected into the image plane by a convex (eye or camera) lens at the origin. It is inconvenient to have to consider inverted images, and it is a commonly used convention in image analysis to set the center of the lens at the origin (0, 0, 0) and to imagine the image plane to be the plane $Z = f$, f being the focal length of the lens; with this simplified geometry (Fig. 16.3B), images in the image plane appear noninverted. Taking a general point in the scene as (X, Y, Z), which appears in the image as (x_1, y_1), perspective projection now gives:

$$(x_1, y_1) = \left(fX/Z,\ fY/Z\right) \tag{16.1}$$

16.3.1 BINOCULAR IMAGES

Fig. 16.4 shows the situation when two lenses are used to obtain a stereo pair of images. In general, the two optical systems do not have parallel optical axes but exhibit a "vergence" (which may be variable, as it is for human eyes), so that they intersect at some point within the scene. Then a general point (X, Y, Z) in the

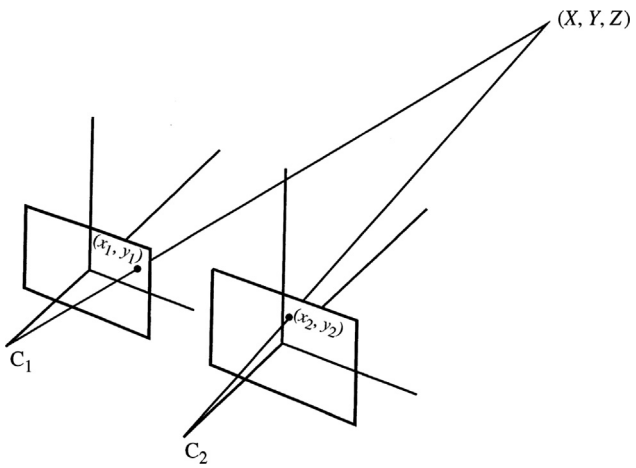

FIGURE 16.4

Stereo imaging using two lenses. The axes of the optical systems are parallel, i.e., there is no "vergence" between the optical axes.

scene has two different pairs of coordinates, (x_1, y_1) and (x_2, y_2), in its two images, which differ both because of the vergence between the optical axes and because the baseline b between the lenses causes relative displacement or "disparity" of the points in the two images.

For simplicity, we now take the vergence to be zero, i.e., the optic axes are parallel. Then, with suitable choice of Z axis on the perpendicular bisector of the baseline b, we obtain two equations:

$$x_1 = (X + b/2)f/Z \tag{16.2}$$

$$x_2 = (X - b/2)f/Z \tag{16.3}$$

so that the disparity is

$$D = x_1 - x_2 = bf/Z \tag{16.4}$$

Rewriting this equation in the form:

$$Z = bf/(x_1 - x_2) \tag{16.5}$$

now permits the depth Z to be calculated. In fact, computation of Z only requires the disparity for a stereo pair of image points to be found and parameters of the optical systems to be known. However, confirming that both points in a stereo pair actually correspond to the same point in the original scene is in general not at all trivial, and much of the computation in stereo vision is devoted to this task. In addition, to obtain good accuracy in the determination of depth, a large baseline b is required: unfortunately, as b is increased the correspondence between the images decreases, so it becomes more difficult to find matching points.

16.3.2 THE CORRESPONDENCE PROBLEM

There are two important approaches to finding pairs of points that match in the two images of a stereo pair. One is that of "light striping" (one form of structured lighting), which encodes the two images so that it is easy to see pairs of corresponding points. If a single vertical stripe is used, for every value of y there is in principle only one light stripe point in each image and so the matching problem is solved. We return to this problem in a later section.

The second important approach is to employ epipolar lines. To understand this approach, imagine that we have located a distinctive point in the first image and that we are marking all possible points in the object field which could have given rise to it. This will mark out a line of points at various depths in the scene and, when viewed in the second image plane, a locus of points can be constructed in that plane. This locus is the *epipolar line* corresponding to the original image point in the alternate image (Fig. 16.5). If we now search along the epipolar line for a similarly distinctive point in the second image, the chance of finding the correct match is significantly enhanced. This method has the advantage not only of cutting down the amount of computation required to find corresponding points, but also of reducing significantly the chance of false alarms. Note that the concept of an epipolar line applies to both images—a point in one image gives an epipolar line in the other image. Note also that in the simple geometry of Fig. 16.4, all epipolar lines are parallel to the x-axis, although this is

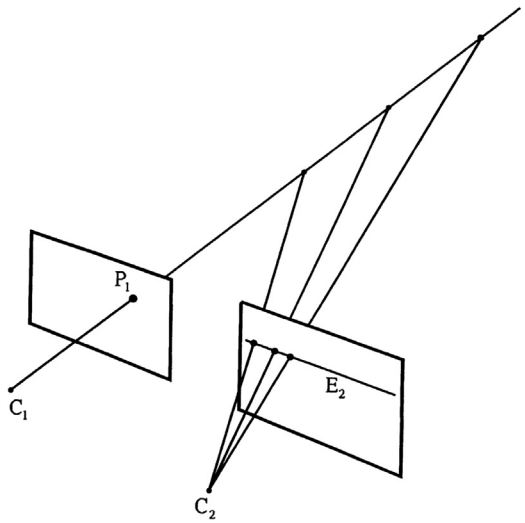

FIGURE 16.5

Geometry of epipolar lines. A point P_1 in one image plane may have arisen from any one of a line of points in the scene and may appear in the alternate image plane at any point on the so-called epipolar line E_2.

not so in general (in fact, the general situation is that all epipolar lines in one image plane pass through the point that is the image of the projection point of the alternate image plane).

The correspondence problem is rendered considerably more difficult by the fact that there will be points in the scene which give rise to points in one image but not in the other. Such points are either occluded in the one image, or else are so distorted as not to give a recognizable match in the two images (e.g., the different background might mask a corner point in one image while permitting it to stand out in the other). Any attempt to match such points can then only lead to false alarms. Thus, it is necessary to search for consistent sets of solutions in the form of continuous object surfaces in the scene. For this reason, iterative "relaxation" schemes are widely used to implement stereo matching.

Broadly speaking, correspondences are sought by two methods: one is the matching of near-vertical edge points in the two images (near-horizontal edge points do not give the required precision); the other is the matching of local intensity patterns using correlation techniques. Correlation is an expensive operation and, in this case, is relatively unreliable—principally because intensity patterns frequently appear significantly foreshortened (i.e., distorted by the effects of perspective) in one or other image and hence are difficult to match reliably. In such cases, the most practical solution is to reduce the baseline; as noted earlier, this has the effect of reducing the accuracy of depth measurement. Further details of these techniques are to be found in Shirai (1987).

Before leaving this topic, we consider in slightly more detail how the above-mentioned problems of visibility arise. Fig. 16.6 shows a situation in which an object is being observed by two cameras giving stereo images. Clearly, much of the object will not be visible in either image because of self-occlusion, whereas some feature points will only be visible in one or the other image. Now consider the order in which the points appear in the two images (Fig. 16.7). The points which are visible appear in the same order as in the scene, and the points which are just going out of sight are those for which the order between the scene and the image is just about to change. Points which provide information about the front surface of the object can thus only bear a simple geometrical relation to each other: in particular, for points not to obscure, or be obscured by, a given point P, they must not lie within a double-ended cone region defined by P and the centers of projection C_1, C_2 of the two cameras: this region is shown shaded in Fig. 16.7. A surface passing through P for which full-depth information can be retrieved must lie entirely within the nonshaded region. (Of course, a new double-ended cone must be considered for each point on the surface being viewed.) Note that the possibility of objects containing holes, or having transparent sections, must not be forgotten (such cases can be detected from differences in the ordering of feature points in the two views—see Fig. 16.7); neither must it be ignored that the foregoing figures represent a single horizontal cross section of an object which can have totally different shapes and depths in different cross-sections.

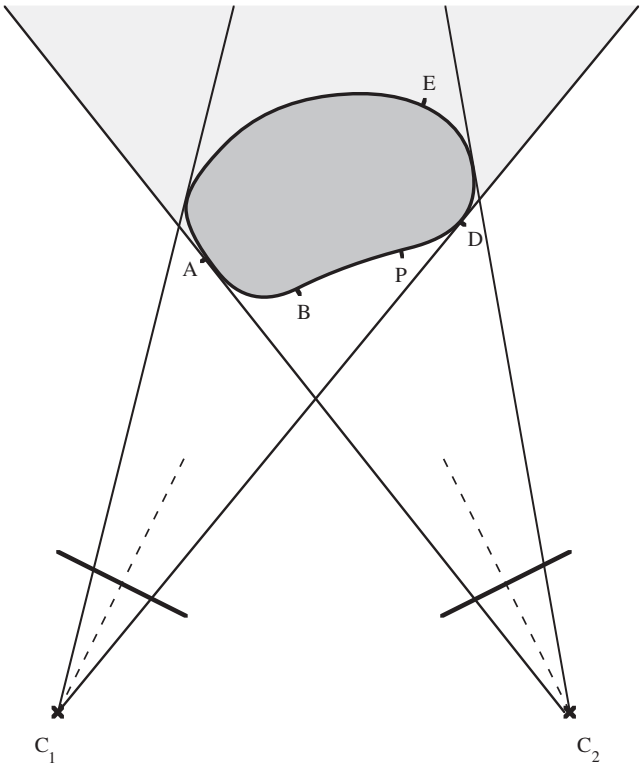

FIGURE 16.6

Visibility of feature points in two stereo views. Here, an object is viewed from two directions. Only feature points which appear in both views are of value for depth estimation. This eliminates all points in the shaded region, such as E, from consideration.

16.4 SHAPE FROM SHADING

It was mentioned in Section 16.2 that it is possible to analyze the pattern of intensities in a single (monocular) image and to deduce the shapes of objects from the shading information. The principle underlying this technique is that of modeling the reflectance of objects in the scene as a function of the angles of incidence i and emergence e of light from their surfaces. In fact, a third angle is also involved, and it is called the "phase" g (Fig. 16.8).

A general model of the situation gives the radiance I (light intensity in the image) in terms of the irradiance E (energy per unit area falling on the surface of the object) and the reflectance R:

$$I(x_1, y_1) = E(x, y, z)R(\mathbf{n}, \mathbf{s}, \mathbf{v}) \tag{16.6}$$

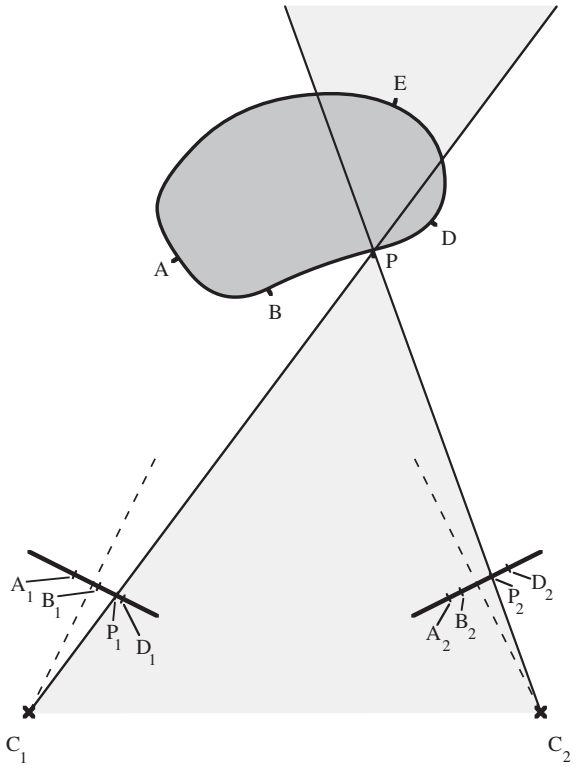

FIGURE 16.7

Ordering of feature points on an object. In the two views of the object shown here, the feature points all appear in the same order A, B, P, D as on the surface of the object. Points for which this would not be valid, such as E, are behind the object and are obscured from view. Relative to a given visible feature P, there is a double-ended cone (shaded) in which feature points must not appear if they are not to obscure the feature under consideration. An exception to these rules might be if the object had a semitransparent window through which an additional feature T were visible: in that case, interpretation would be facilitated by noting that the orderings of the features seen in the two views were different—e.g., A_1, T_1, B_1, P_1, D_1 and A_2, B_2, T_2, P_2, D_2.

It is well known that a number of matt surfaces approximate reasonably well to an ideal Lambertian surface whose reflectance function depends only on the angle of incidence i—that is, the angles of emergence and phase are immaterial:

$$I = (1/\pi)E \cos i \tag{16.7}$$

For the present purpose E is regarded as a constant and is combined with other constants for the camera and the optical system (including, e.g., the

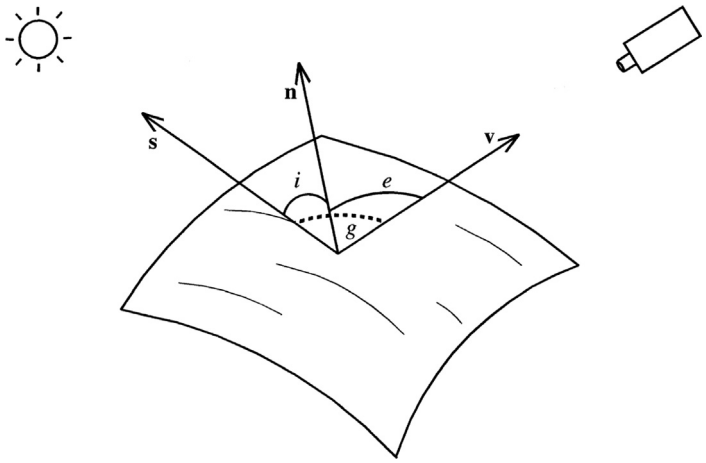

FIGURE 16.8

Geometry of reflection. An incident ray from source direction **s** is reflected along the viewer direction **v** by an element of the surface whose local normal direction is **n**; *i*, *e*, and *g* are defined respectively as the incident, emergent, and phase angles.

f-number). In this way, a normalized reflectance is obtained, which in this case is

$$\begin{aligned} R &= R_0 \cos i = R_0 \mathbf{s}\cdot\mathbf{n} \\ &= \frac{R_0(1 + pp_s + qq_s)}{(1+p^2+q^2)^{1/2}(1+p_s^2+q_s^2)^{1/2}} \end{aligned} \tag{16.8}$$

where we have used the standard convention of writing orientations in 3-D in terms of p and q values. These are not direction cosines but correspond to the coordinates of the point (p, q, l) at which a particular direction vector from the origin meets the plane $z = 1$: hence, they need suitable normalization, as in the above equation.

The above equation gives a reflectance map in gradient (p, q) space. We now temporarily set the absolute reflectance value R_0 equal to unity. The reflectance map can be drawn as a set of contours of equal brightness, starting with a point having $R = 1$ at $\mathbf{s} = \mathbf{n}$, and going down to zero for **n** perpendicular to **s**. When $\mathbf{s} = \mathbf{v}$, so that the light source is along the viewing direction (here taken to be the direction $p = q = 0$), zero brightness occurs only for infinite distances on the reflectance map $[(p^2+q^2)^{1/2}$ approaching infinity] (Fig. 16.9A). In a more general case, when $\mathbf{s} \neq \mathbf{v}$, zero brightness occurs along a straight line in gradient space (Fig. 16.9B). To find the exact shapes of the contours, we can set R at a constant value a, which results in

$$a(1+p^2+q^2)^{1/2}(1+p_s^2+q_s^2)^{1/2} = 1 + pp_s + qq_s \tag{16.9}$$

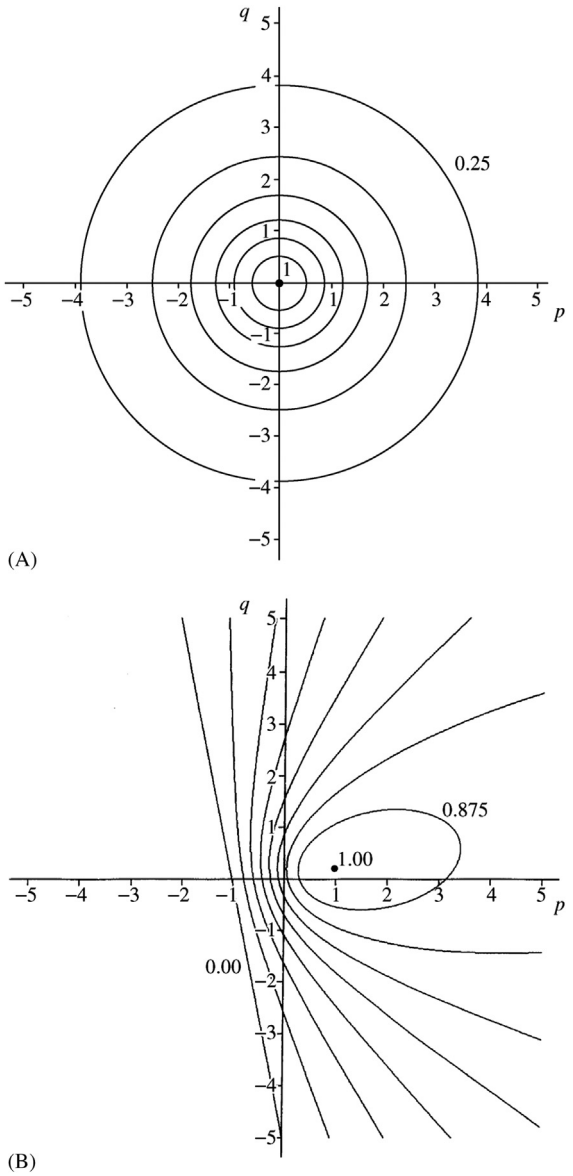

FIGURE 16.9

Reflectance maps for Lambertian surfaces: (A) contours of constant intensity plotted in gradient (p, q) space for the case where the source direction **s** (marked by a black dot) is along the viewing direction **v** (0, 0) (the contours are taken in steps of 0.125 between the values shown); (B) the contours that arise where the source direction (p_s,q_s) is at a point (marked by a black dot) in the positive quadrant of (p, q) space: note that there is a well-defined region, bounded by the straight line $1 + pp_s + qq_s = 0$, for which the intensity is zero (the contours are again taken in steps of 0.125).

Squaring this equation clearly gives a quadratic in p and q, which could be simplified by a suitable change of axes. Thus, the contours must be curves of conic section, namely circles, ellipses, parabolas, hyperbolas, lines, or points (the case of a point arises only when $a = 1$, when we get $p = p_s$, $q = q_s$; and that of a line only if $a = 0$, when we get the equation $1 + pp_s + qq_s = 0$: both of these solutions were implied above).

Unfortunately, object reflectances are not all Lambertian and an obvious exception is for surfaces that approximate to pure specular reflection. In that case $e = i$ and $g = i + e$ (**s**, **n**, **v** are coplanar); the only nonzero reflectance position in gradient space is the point representing the bisector of the angle between the source direction **s** (p, q) and the viewing direction **v** (0, 0)—that is, **n** is along **s** + **v**—and very approximately:

$$p \approx p_s/2 \tag{16.10}$$

$$q \approx q_s/2 \tag{16.11}$$

For less perfect specularity, a peak is obtained around this position. A good approximation to the reflectance of many real surfaces is obtained by modeling them as basically Lambertian but with a strong additional reflectance near the specular reflectance position. Using the Phong (1975) model for the latter component gives

$$R = R_0 \cos i + R_1 \cos^m \theta \tag{16.12}$$

θ being the angle between the actual emergence direction and the ideal specular reflectance direction.

The resulting contours now have two centers around which to peak: the first is the ideal specular reflection direction ($p \approx p_s/2$, $q \approx q_s/2$), and the second is that of the source direction ($p = p_s$, $q = q_s$). When objects are at all shiny—such as metal, plastic, liquid, or even wood surfaces—the specular peak is quite sharp and rather intense: casual observation may not even indicate the presence of another peak as Lambertian reflection is so diffuse (Fig. 16.10). In other cases, the specular peak can broaden and become more diffuse: hence it may merge with the Lambertian peak and effectively disappear.

Some remarks should be made about the Phong model employed above. First, it is adapted to different materials by adjusting the values of R_0, R_1, and m. Phong remarks that R_1 typically lies between 10% and 80%, whereas m is in the range 1−10. However, Rogers (1985) indicates that m may be as high as 50. Note that there is no physical significance in these numbers—the model is simply a phenomenological one. This being so, care should be taken to prevent the $\cos^m\theta$ term from contributing to reflectance estimates when $|\theta| > 90°$. The Phong model is reasonably accurate but has been improved by Cook and Torrance (1982). This is important in computer graphics applications but the improvement is difficult to apply in computer vision, because of lack of data concerning the reflectances of real objects and because of variability in the current state (cleanliness, degree of

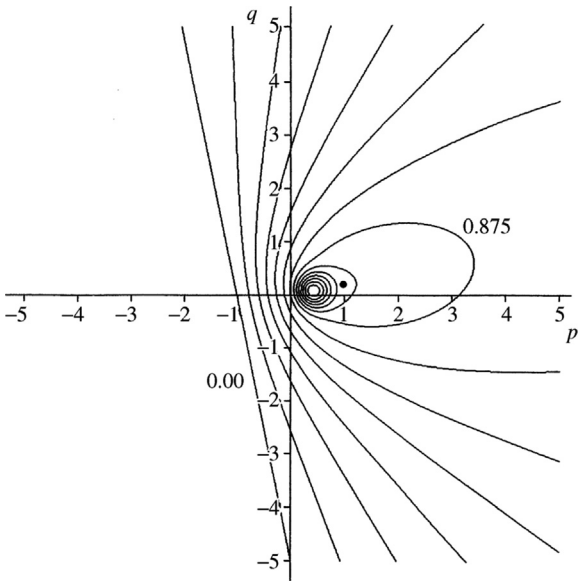

FIGURE 16.10

Reflectance map for a non-Lambertian surface: a modified form of Fig. 16.9B for the case where the surface has a marked specular component ($R_0 = 1.0$, $R_1 = 0.8$): note that the specular peak can have very high intensity (much greater than the maximum value of unity for the Lambertian component). In this case, the specular component is modeled with a $\cos^8\theta$ variation (the contours are again taken in steps of 0.125).

polish, etc.) of a given surface. However, the method of photometric stereo gives some possibility of overcoming these problems.

16.5 PHOTOMETRIC STEREO

Photometric stereo is a form of structured lighting that increases the information available from surface reflectance variations. Basically, instead of taking a single monocular image of a scene illuminated from a single source, several images are taken, from the same vantage point, with the scene illuminated in turn by separate light sources. These light sources are ideally point sources some distance away in various directions, so that there is in each case a well-defined light source direction from which to measure surface orientation.

The basic idea of photometric stereo is that of cutting down the number of possible positions in gradient space for a given point on the surface of an object. It has already been seen that, for known absolute reflectance R_0, a constant brightness in one image permits the surface orientation to be limited to a curve of conic

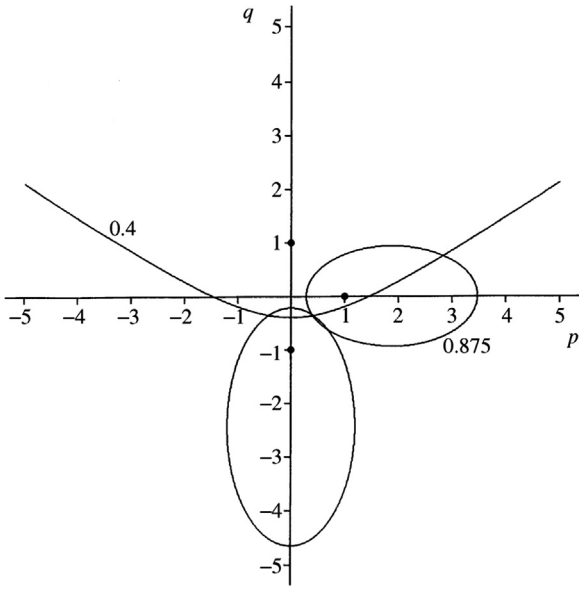

FIGURE 16.11

Obtaining a unique surface orientation by photometric stereo. Three contours of constant intensity arise for different light sources of equal strength: all three contours pass through a single point in (p,q) space and result in a unique solution for the local gradient.

cross section in gradient space. This would also be true for a second such image, the curve being a new one if the illuminating source is different. In general, two such conic curves meet at two points, so there is now only a single ambiguity in the gradient of the surface at any given point in the image. To resolve this ambiguity, a third source of illumination can be employed (this must not be in the plane containing the first two and the surface point being examined), and the third image gives another curve in gradient space which should pass through the appropriate crossing point of the first two curves (Fig. 16.11). If a third source of illumination cannot be used, it is sometimes possible to arrange that the inclination of each of the sources is so high that $(p^2+q^2)^{1/2}$ on the surface is always lower than $(p_s^2+q_s^2)^{1/2}$ for each of the sources, so that only one interpretation of the data is possible. This method is prone to difficulty, however, as it means that parts of the surface could be in shadow, thereby preventing the gradient for these parts of the surface from being measured. Another possibility is to assume that the surface is reasonably smooth, so that p and q vary continuously over it. This itself ensures that ambiguities are resolved over most of the surface.

However, there are other advantages to be gained from using more than two sources of illumination. One is that information on the absolute surface reflectance can be obtained. Another is that the assumption of a Lambertian surface can

be tested. Thus, three sources of illumination ensure that the remaining ambiguity is resolved *and* permit absolute reflectivity to be measured: this is obvious as if the three contours in gradient space do not pass through the same point, then the absolute reflectivity cannot be unity, so corresponding contours should be sought which do pass through the same point. In practice the calculation is normally carried out by defining a set of nine matrix components of irradiance, s_{ij} being the jth component of light source vector $\mathbf{s}_i$. Then, in matrix notation:

$$\mathbf{E} = R_0 S \mathbf{n} \tag{16.13}$$

where

$$\mathbf{E} = (E_1,\ E_2,\ E_3)^{\mathrm{T}} \tag{16.14}$$

and

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix} \tag{16.15}$$

Provided that the three vectors $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$ are not coplanar, so that S is not a singular matrix, R_0 and $\mathbf{n}$ can now be determined from the formulae:

$$R_0 = |S^{-1}\mathbf{E}| \tag{16.16}$$

$$\mathbf{n} = \frac{S^{-1}\mathbf{E}}{R_0} \tag{16.17}$$

An interesting special case arises if the three source directions are mutually perpendicular; taking them to be aligned along the respective major axes directions, S is now the unit matrix, so that:

$$R_0 = (E_1^2 + E_2^2 + E_3^2)^{1/2} \tag{16.18}$$

and

$$\mathbf{n} = (E_1, E_2, E_3)^{\mathrm{T}} / R_0 \tag{16.19}$$

If four or more images are obtained using further illumination sources, more information can be obtained: for example, the coefficient of specular reflectance, R_1. In practice, this coefficient varies somewhat randomly with the cleanness of the surface and it may not be relevant to determine it accurately. More probably, it will be sufficient to check whether significant specularity is present, so that the corresponding region of the surface can be ignored for absolute reflectance calculations. Nonetheless, finding the specularity peak can itself give important surface orientation information, as will be clear from the previous section. Note that, although the information from several illumination sources should ideally be collated using least-squares analysis, this method requires significant computation. Hence, it seems better to use the images resulting from further illumination sources as confirmatory—or, instead, to select the three that exhibit the

least evidence of specularity as giving the most reliable information on local surface orientation.

16.6 THE ASSUMPTION OF SURFACE SMOOTHNESS

It was hinted above that the assumption of a reasonably smooth surface permits ambiguities to be removed in situations where there are two illuminating sources. In fact, this method can be used to help analyze the brightness map even for situations where a single source is employed: indeed, the fact that the eye can perform this feat of interpretation indicates that it should be possible to find computer methods for achieving it. Much research has been carried out on this topic and a set of methods is available, although the calculations are complex, iterative, computation intensive procedures. For this reason, they are not studied in depth here: the reader is referred to the volume by Horn (1986) for detailed information on this topic. However, one or two remarks are in order.

First, consider the representation to be employed for this type of analysis. In fact, normal gradient (p, q) space is not very appropriate for the purpose. In particular, it is necessary to average gradient (i.e., the **n** values) locally within the image; however, (p, q)-space is not "linear," in that a simple average of (p, q) values within a window would give biased results. It turns out that a conformal representation of gradient (i.e., one which preserves small shapes) is closer to the ideal, in that the distances between points in such a representation provide better approximations to the relative orientations of surface normals: averaging in such a representation gives reasonably accurate results. The required representation is obtained by a stereographic projection, which maps the unit (Gaussian) sphere onto a plane $(z = 1)$ through its north pole but this time using as a projection point not its center but its south pole. This projection has the additional advantage that it projects all possible orientations of a surface onto the plane, not merely those from the northern hemisphere. Hence, backlit objects can be represented conveniently in the same map as used for frontlit objects.

Second, the relaxation methods used to estimate surface orientation have to be provided with accurate boundary conditions: in principle, the more correct the orientations that are presented initially to such procedures, the more quickly and accurately the iterations proceed. There are normally two sets of boundary condition that can be applied in such programs. One is the set of positions in the image where the surface normal is perpendicular to the viewing direction. The other is the set of positions in the image where the surface normal is perpendicular to the direction of illumination: this set of positions corresponds to the set of shadow edges (Fig. 16.12). Careful analysis of the image must be undertaken to find each set of positions, but once they have been located they provide valuable cues for unlocking the information content of the monocular image, and mapping out surfaces in detail.

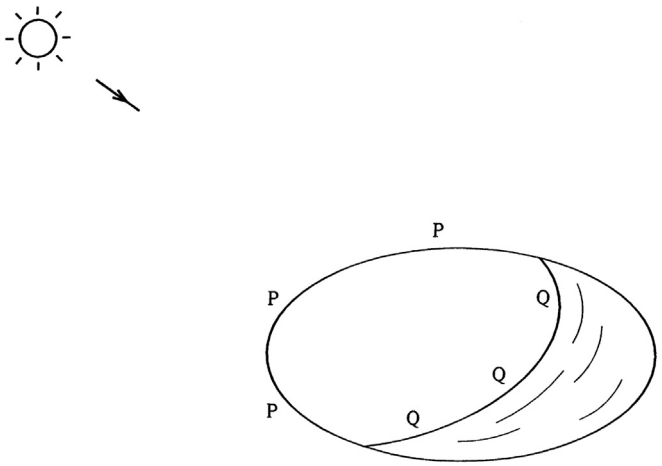

FIGURE 16.12

The two types of boundary condition that can be used in shape-from-shading computations of surface orientation: (1) positions P where the surface normal is perpendicular to the viewing direction; (2) positions Q where the surface normal is perpendicular to the direction of illumination (i.e., shadow boundaries).

Finally, all shapes from shading techniques provide information which initially takes the form of surface orientation maps. Dimensions are not obtainable directly but these can be computed by integration across the image from known starting points. In practice, this tends to mean that absolute dimensions are unknown and that dimensional maps are obtainable only if the size of an object is given or if its depth within the scene is known.

16.7 SHAPE FROM TEXTURE

Texture can be very helpful to the human eye in permitting depth to be perceived. Although textured patterns can be very complex, even the simplest textural elements can carry depth information. Ohta et al. (1981) showed how circular patches on a flat surface viewed more and more obliquely in the distance become first elliptical and then progressively flatter and flatter. At infinite distance, on the horizon line (here defined as the line at infinity in the given plane), they would clearly become very short line segments. To disentangle such textured images sufficiently to deduce depths within the scene, it is first necessary to find the horizon line reliably. This is achieved by taking all pairs of texture elements and deducing from their areas where the horizon line would have to be. To proceed, we make use of the rule:

$$d_1^3/d_2^3 = A_1/A_2 \tag{16.20}$$

which applies as circles at various depths would give a square law, although the progressive eccentricity also reduces the area linearly in proportion to the depth. This information is accumulated in a separate image space and a line is then fitted to these data: false alarms are eliminated automatically by this Hough-based procedure. At this stage the original data—the ellipse areas—provide direct information on depth, although some averaging is required to obtain accurate results. Although this type of method has been demonstrated in certain instances, it is in practice highly restricted unless very considerable amounts of computation are performed. Hence, it is doubtful whether it can be of general practical use in machine vision applications.

16.8 USE OF STRUCTURED LIGHTING

Structured lighting has already been considered briefly in Section 16.2 as an alternative to stereo for mapping out depth in scenes. Basically, a pattern of light stripes, or other arrangement of light spots or grids, is projected onto the object field. Then these patterns are enhanced in a (generally) single monocular image and analyzed to extract the depth information. To obtain the maximum information, the light pattern must be close-knit and the received images must be of very high resolution. When shapes are at all complex, the lines can in places appear so close together that they are unresolvable. It then becomes necessary to separate the elements in the projected pattern, trading resolution and accuracy for reliability of interpretation. Even so, if parts of the objects are along the line of sight, the lines can merge together and even cross back and fore, so unambiguous interpretation is never assured. In fact, this is part of a larger problem, in that parts of the object will be obscured from the projected pattern by occluding bodies or by self-occlusion: the method has this feature in common with the shape from shading technique and with stereo vision, which relies on *both* cameras being able to view various parts of the objects simultaneously. Hence, the structured light approach is subject to similar restrictions to those found for other methods of 3-D vision and is not a panacea. Nevertheless, it is a useful technique that is generally simple to set up so as to acquire specific 3-D information which can enable a computer to start the process of cueing into complex images.

Light spots provide perhaps the most obvious form of structured light. However, they are restricted because for each spot an analysis has to be performed to determine which spot is being viewed: connected lines, in contrast, carry a large amount of coding information with them so that ambiguities are less likely to arise. Grids of lines carry even more coding information but do not necessarily give any more depth information. Indeed, if a pattern of light stripes can be projected (for example) from the left of the camera so that they are parallel to the y-axis in the observed image, then there is no point in projecting another set of lines parallel to the x-axis, as these merely replicate information that is already

available from the rows of pixels in the image—all the depth information is carried by the vertical lines and their horizontal displacements in the image. This analysis assumes that the camera and projected beams are carefully aligned, and that no perspective or other distortions are present. In fact, most practical structured lighting systems in current use employ light stripe patterns rather than spot patterns or full-grid patterns.

This section ends with an analysis of the situations that can arise when a single stripe is incident on objects as simple as rectangular blocks. Fig. 16.13 shows three types of structure in observed stripes: (1) the effect of a sharp angle being encountered; (2) the effect of "jump edges" at which light stripes jump horizontally and vertically at the same time; and (3) the effect of discontinuous edges at which light stripes jump horizontally but not vertically. The reasons for these circumstances will be obvious from Fig. 16.13. Basically, the problem to be tackled with jump and discontinuous edges is to find whether a given stripe end marks an occluding edge or an occluded edge. The importance of this distinction is that occluding edges mark actual edges of the object being observed, whereas occluded edges may be merely edges of shadow regions and are then not *directly* significant (more precisely, they involve interactions of light with two objects rather than with one and are therefore more complex to interpret). A simple rule is that, if stripes are projected from the left, the left-hand component of a discontinuous edge will be the occluding edge and the right-hand component will be the

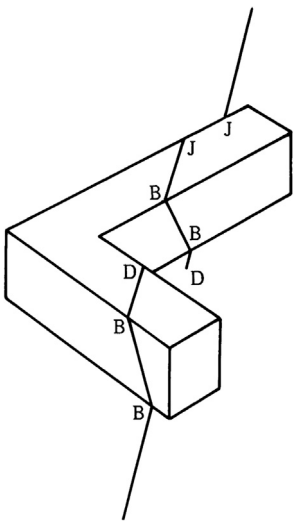

FIGURE 16.13

Three of the structures that are observed when a light stripe is incident on even quite simple shapes: bends (B), jumps (J), and discontinuities (D).

occluded edge. Angle edges are located by applying a Laplacian type of operator which detects the change in orientation of the light stripe.

The ideas outlined above correspond to possible 1-D operators that interpret light stripe information to locate nonvertical edges of objects. The method provides no direct information concerning vertical edges. To obtain such information, it is necessary to analyze the information from sets of light stripes. For this purpose 2-D edge operators are required, which collect sufficient data from at least two or three adjacent light stripes. Further details are beyond the scope of this chapter.

Overall, light stripes provide a very useful means of recognizing planes forming the faces of polyhedra and other types of manufactured object. The characteristic sets of parallel lines can be found and demarcated relatively easily, and the fact that the lines usually give rather strong signals means that line-tracking techniques can be applied and that algorithms can operate quite rapidly. However, whole-scene interpretation, including inferring the presence and relative positions of different objects, remains a more complex task, as will be seen below.

16.9 THREE-DIMENSIONAL OBJECT RECOGNITION SCHEMES

The methods described so far in this chapter employ various means for finding depth at all places in a scene and are hence able to map out 3-D surfaces in a fair amount of detail. However, they do not give any clue as to what these surfaces represent. In some situations, it may be clear that certain planar surfaces are parts of the background, e.g., the floor and the walls of a room, but in general individual objects will not be inherently identifiable. Indeed, objects tend to merge with each other and with the background, so specific methods are needed to segment the 3-D space map and finally recognize the objects, giving detailed information on their positions and orientations. [A 3-D space map may be defined as an imagined 3-D map showing, without interpretation, the surfaces of all objects in the scene and incorporating all the information from depth or range images. Note that it will generally include only the front surfaces of objects seen from the vantage point of the camera.]

Before proceeding to study this problem, notice that further general processing can be carried out to analyze the 3-D shapes. Agin and Binford (1976) and others have developed techniques for likening 3-D shapes to "generalized cylinders", these being like normal (right circular) cylinders but with additional degrees of freedom so that the axes can bend and the cross-sections can vary, both in size and in detailed shape: even an animal like a sheep can be likened to a distorted cylinder. On the whole, this approach is elegant but may not be well adapted to describe many industrial objects, and it is therefore not pursued further here. A simpler approach may be to model the 3-D surfaces as planar, quadratic, cubic, and quartic surfaces, and then to try to understand these model surfaces in terms of what is known about existing objects. This approach was adopted by Hall et al.

(1982) and was found to be viable, at least for certain quite simple objects such as cups. Shirai (1987) has taken the approach even further so that a whole range of objects can be found and identified in quite complex indoor scenes.

We next consider what we are trying to achieve regarding recognition. First, can recognition be carried out *directly* on the mapped out 3-D surfaces, just as it could for the 2-D images of earlier chapters? Second, if we can bypass the 3-D modeling process, and still recognize objects, might it not be possible to save even more computation and omit the stage of mapping out 3-D surfaces, instead identifying 3-D objects directly in 2-D images? It might even be possible to locate 3-D objects from a single 2-D image.

Consider the first of these problems. When we studied 2-D recognition, many instances were found where the Hough Transform (HT) approach was of great help. It turned out to give trouble in more complex cases, particularly when attempts were made to find objects where there were more than two or at most three degrees of freedom. Here, however, we have situations where objects normally have six degrees of freedom—three degrees of freedom for translation and another three for rotation. This doubling of the number of free parameters on going from 2-D to 3-D makes the situation far worse, as the search space is proportional in size not to the number of degrees of freedom, but to its exponent: for example, if each degree of freedom in translation or rotation can have 256 values, the number of possible locations in parameter space changes from 256^3 in 2-D to 256^6 in 3-D. This will be seen to have a very profound effect on object location schemes and tends to make the HT technique difficult to implement. In the next section, we study an interesting approach to the 3-D recognition problem, which uses a subtle combination of 2-D and 3-D techniques.

16.10 HORAUD'S JUNCTION ORIENTATION TECHNIQUE

This and related techniques are sometimes referred to as "shape from angle."

Horaud's (1987) technique is special in that it uses as its starting point 2-D images of 3-D scenes and "backprojects" them into the scene, with the aim of making interpretations in 3-D rather than 2-D frames of reference. This has the initial effect of increasing mathematical complexity, although in the end useful, more accurate results emerge.

Initially, the boundaries of planar surfaces on objects are backprojected. Each boundary line is thus transformed into an "interpretation plane" defined by the center of the camera projection system and the boundary line in the image plane: clearly, the interpretation plane must contain the line that originally projected into the boundary line in the image. Similarly, angles between boundary lines in the image are backprojected into two interpretation planes, which must contain the original two object lines. Finally, junctions between three boundary lines are backprojected into three interpretation planes which must contain a corner in the

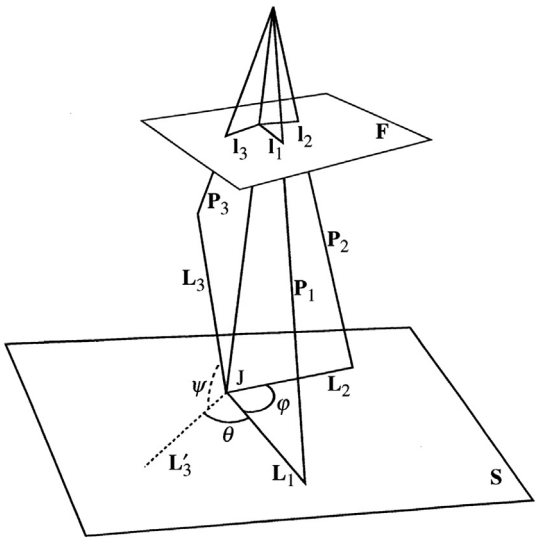

FIGURE 16.14

Geometry for backprojection from junctions: a junction of three lines in an image may be backprojected into three planes, from which the orientation in space of the original corner J may be deduced.

space map (Fig. 16.14). The paper focuses on the backprojection of junctions and shows how measurements of the junction angles in the image relate to those of the original corner; it also shows how the space orientation of the corner can be computed. In fact, it is interesting that the orientation of an object in 3-D can in general be deduced from the appearance of just one of its corners in a single image. This is a powerful result and in principle permits objects to be recognized and located from extremely sparse data.

To understand the method, the mathematics first needs to be set up with some care. Assumes that lines $\mathbf{L}_1$, $\mathbf{L}_2$, $\mathbf{L}_3$ meet at a junction in an object and appear as lines $\mathbf{l}_1$, $\mathbf{l}_2$, $\mathbf{l}_3$ in the image (Fig. 16.14). Take respective interpretation planes containing the three lines and label them by unit vectors $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ along their normals, so that:

$$\mathbf{P}_1 \cdot \mathbf{L}_1 = 0 \tag{16.21}$$

$$\mathbf{P}_2 \cdot \mathbf{L}_2 = 0 \tag{16.22}$$

$$\mathbf{P}_3 \cdot \mathbf{L}_3 = 0 \tag{16.23}$$

In addition, take the space plane containing $\mathbf{L}_1$ and $\mathbf{L}_2$, and label it by a unit vector $\mathbf{S}$ along its normal, so that

$$\mathbf{S}_1 \cdot \mathbf{L}_1 = 0 \tag{16.24}$$

$$\mathbf{S}_2 \cdot \mathbf{L}_2 = 0 \tag{16.25}$$

As $\mathbf{L}_1$ is perpendicular to $\mathbf{S}$ and to $\mathbf{P}_1$, and $\mathbf{L}_2$ is perpendicular to $\mathbf{S}$ and to $\mathbf{P}_2$, it is found that

$$\mathbf{L}_1 = \mathbf{S} \times \mathbf{P}_1 \tag{16.26}$$

$$\mathbf{L}_2 = \mathbf{S} \times \mathbf{P}_2 \tag{16.27}$$

Note that $\mathbf{S}$ is not in general perpendicular to $\mathbf{P}_1$ and $\mathbf{P}_2$, so $\mathbf{L}_1$ and $\mathbf{L}_2$ are not in general unit vectors. Defining φ as the angle between $\mathbf{L}_1$ and $\mathbf{L}_2$, we now have

$$\mathbf{L}_1 \cdot \mathbf{L}_2 = L_1 L_2 \cos \varphi \tag{16.28}$$

which can be reexpressed in the form:

$$(\mathbf{S} \times \mathbf{P}_1) \cdot (\mathbf{S} \times \mathbf{P}_2) = |\mathbf{S} \times \mathbf{P}_1| \ |\mathbf{S} \times \mathbf{P}_2| \cos \varphi \tag{16.29}$$

Next, we need to consider the junction between $\mathbf{L}_1$, $\mathbf{L}_2$, $\mathbf{L}_3$. To proceed, it is necessary to specify the relative orientations in space of the three lines. θ is the angle between $\mathbf{L}_1$ and the projection $\mathbf{L}'_3$ of $\mathbf{L}_3$ on plane $\mathbf{S}$, whereas ψ is the angle between $\mathbf{L}'_3$ and $\mathbf{L}_3$ (Fig. 16.14). Thus, the structure of the junction J is described completely by the three angles φ, θ, ψ. $\mathbf{L}_3$ can now be found in terms of other quantities:

$$\mathbf{L}_3 = \mathbf{S} \sin \psi + \mathbf{L}_1 \cos \theta \cos \psi + (\mathbf{S} \times \mathbf{L}_1) \sin \theta \cos \psi \tag{16.30}$$

Applying Eq. (16.23), we find

$$\mathbf{S} \cdot \mathbf{P}_3 \sin \psi + \mathbf{L}_1 \cdot \mathbf{P}_3 \cos \theta \cos \psi + (\mathbf{S} \times \mathbf{L}_1) \cdot \mathbf{P}_3 \sin \theta \cos \psi = 0 \tag{16.31}$$

Substituting for $\mathbf{L}_1$ from Eq. (16.26), and simplifying, we finally obtain

$$\begin{aligned}(\mathbf{S} \cdot \mathbf{P}_3)|\mathbf{S} \times \mathbf{P}_1| \sin \psi + \mathbf{S} \cdot (\mathbf{P}_1 \times \mathbf{P}_3) \cos \theta \cos \psi \\ + (\mathbf{S} \cdot \mathbf{P}_1)(\mathbf{S} \cdot \mathbf{P}_3) \sin \theta \cos \psi = (\mathbf{P}_1 \cdot \mathbf{P}_3) \sin \theta \cos \psi\end{aligned} \tag{16.32}$$

Eqs. (16.31) and (16.34) now exclude the unknown vectors $\mathbf{L}_1$, $\mathbf{L}_2$, $\mathbf{L}_3$ but they retain $\mathbf{S}$, $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ and the three angles φ, θ, ψ. $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ are known from the image geometry, and the angles φ, θ, ψ are presumed to be known from the object geometry; in addition, only two components (α, β) of the unit vector $\mathbf{S}$ are independent, so the two equations should be sufficient to determine the orientation of the space plane $\mathbf{S}$. Unfortunately, the two equations are highly nonlinear, and it is necessary to solve them numerically. Horaud (1987) achieved this by reexpressing the formulae in the forms:

$$\cos \varphi = f(\alpha, \beta) \tag{16.33}$$

$$\begin{aligned}\sin \theta \cos \psi = g_1(\alpha, \beta) \sin \psi + g_2(\alpha, \beta) \cos \theta \cos \psi \\ + g_3(\alpha, \beta) \sin \theta \cos \psi\end{aligned} \tag{16.34}$$

For each image junction, $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ are known and it is possible to evaluate f, g_1, g_2, g_3. Then, assuming a particular interpretation of the junction, values are assigned to φ, θ, ψ and curves giving the relation between α and β are plotted for each equation. Possible orientations for the space plane $\mathbf{S}$ are then given by

positions in (α, β) space where the curves cross. Horaud showed that, in general, 0, 1, or 2 solutions are possible: the case of no solutions corresponds to trying to make an impossible match between a corner and an image junction when totally the wrong angles φ, θ, ψ are assumed; one solution is the normal situation; and two solutions arise in the interesting special case when orthographic or near-orthographic projection permits perceptual reversals—that is, a convex corner is interpreted as a concave corner or vice versa. In fact, under orthographic projection the image data from a single corner are insufficient, taken on their own, to give a unique interpretation: in this situation, even the human visual system makes mistakes—as in the case of the well-known Necker cube illusion (see Chapter 17: Tackling the Perspective n-Point Problem). However, when such cases arise in practical situations, it may be better to take the convex rather than the concave corner interpretation as a working assumption, as it has slightly greater likelihood of being correct.

Horaud has shown that such ambiguities are frequently resolved if the space plane orientation is estimated simultaneously for all the junctions bordering the object face in question, by plotting the α and β values for all such junctions on the same α, β graph. For example, with a cube face on which there are three such junctions, nine curves are coincident at the correct solution, and there are nine points where only two curves cross, indicating false solutions. On the other hand, if the same cube is viewed under conditions approximating very closely to orthographic projection, two solutions with nine coincident curves appear and the situation remains unresolved, as before.

Overall, this technique is important in showing that although lines and angles individually lead to virtually unlimited numbers of possible interpretations of 3-D scenes, junctions lead individually to at most two solutions and any remaining ambiguity can normally be eliminated if junctions on the same face are considered together. As has been seen, the exception to this rule occurs when projection is accurately orthographic, although this is a situation that can often be avoided in practice.

So far, we have considered only how a given hypothesis about the scene may be tested: nothing has been said about how assignments of the angles φ, θ, ψ are made to the observed junctions. Horaud's paper discussed this aspect of the work in some depth. In general, the approach is to use a depth-first search technique in which a match is "grown" from the initial most promising junction assignment. In fact, considerable preprocessing of sample data is carried out to find how to rank image features for their utility during depth-first search interpretation. The idea is to order possible alternatives such as linear or circular arcs, convex or concave junctions, short or long lines, etc. In this way, the tree search becomes more planned and efficient at run time. Generally, the more frequently occurring types of feature should be weighted down in favor of the rarer types of feature, for greater search efficiency. In addition, remember that hypothesis generation is relatively expensive in that it demands a stage of backprojection, as described above. Ideally, this stage need be employed only once for each object (in the case that

only a single corner is, initially, considered). Subsequent stages of processing then involve hypothesis verification in which other features of the object are predicted and their presence sought in the image: if found, they are used to refine the existing match; if the match at any stage becomes worse, then the algorithm backtracks and eliminates one or more features and proceeds with other ones. This process is unavoidable, as more than one image feature may be present near a predicted feature.

One of the factors that has been found to make the method converge quickly is the use of grouped rather than individual features, as this tends to decrease the combinatorial explosion in the size of the search: in the present context, this means that attempts should be made to match first all junctions or angles bordering a given object face, and further that a face should be selected that has the greatest number of matchable features around it.

In summary, this approach is successful, as it backprojects from the image and then uses geometrical constraints and heuristic assumptions for matching in 3-D space. It is suitable for matching objects that possess planar faces and straight-line boundaries, hence giving angle and junction features. However, extending the backprojection technique to situations where object faces are curved and have curved boundaries could be significantly more difficult.

16.11 AN IMPORTANT PARADIGM—LOCATION OF INDUSTRIAL PARTS

In this section, we consider the location of a common class of industrial part: this constitutes an important example that has to be solved in one way or another. Here, we go along with the Bolles and Horaud (1986) approach as it leads to sensible solutions and embodies a number of useful didactic lessons. The method starts with a depth map of the scene (obtained in this case using structured lighting).

Fig. 16.15 shows in simplified form the type of industrial part being sought in the images. In typical scenes, several of these parts may appear jumbled on a worktable, with perhaps three or four being piled on top of each other in some places. In such cases, it is vital that the matching scheme be highly robust if most of the parts are to be found, since even when a part is unoccluded, it appears against a highly cluttered and confusing background. However, the parts themselves have reasonably simple shapes and possess certain salient features. In the particular problem cited, each has a cylindrical base with a concentric cylindrical head and also a planar shelf attached symmetrically to the base. To locate such objects, it is natural to attempt to search for circular and straight dihedral edges. In addition, because of the type of data being used, it is useful to search for straight tangential edges, which appear where the sides of curved cylinders are viewed obliquely.

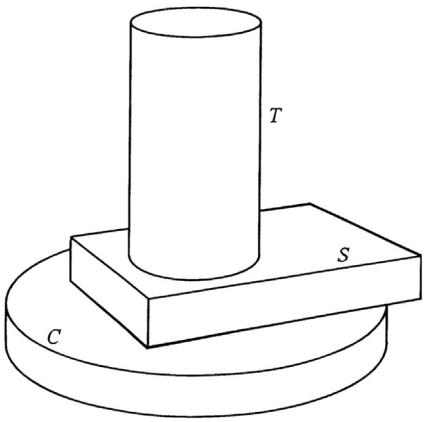

FIGURE 16.15

The essential features of the industrial components located by the 3-D PO system of Bolles and Horaud (1986). S, C, and T indicate, respectively, straight and circular dihedral edges and straight tangential edges, all of which are searched for by the system.

In general, circular dihedral edges appear elliptical, and parameters for five of the six degrees of freedom of the part can be determined by analyzing these edges. The parameter that cannot be determined in this way corresponds to rotation about the axis of symmetry of the cylinder.

Straight dihedral edges also permit five free parameters to be determined, as location of one plane eliminates three degrees of freedom and location of an adjacent plane eliminates a further two degrees of freedom. The parameter that remains undetermined is that of linear motion along the direction of the edge. However, there is also a further ambiguity in that the part may appear either way around on the dihedral edge.

Straight tangential edges determine only four free parameters, as the part is free to rotate about the axis of the cylinder and can also move along the tangential edge. Note that these edges are the most difficult to locate accurately, as range data are subject to greater levels of noise as surfaces curve away from the sensor.

All three of these types of edge are planar. They also provide useful additional information that can help to identify where they are on a part. For example, straight and curved dihedral edges both provide information on the size of the included angle, and the curved edges also give radius values. In fact, curved dihedral edges provide significantly more parametric information about a part than either of the other two types of edge, and therefore they are of most use to form initial hypotheses about the pose (position and orientation) of a part. Having found such an edge, it is necessary to try out various hypotheses about which edge it is, for example, by searching for other circular dihedral edges at specific relative positions: this is a vital hypothesis verification step. Next, the problem of

how to determine the remaining free parameter is solved by searching for the linear straight dihedral edge features from the planar shelf on the part.

At this stage, hypothesis generation is complete and the part is essentially found, but hypothesis verification is required (1) to confirm that the part is genuine and not an accidental grouping of independent features in the image, (2) to refine the pose estimate, and (3) to determine the "configuration" of the part, i.e., to what extent, it is buried under other parts (making it difficult for a robot to pick it up). When the most accurate pose has been obtained, the overall degree of fit can be considered and the hypothesis rejected if some relevant criterion is not met.

In common with other researchers (Faugeras and Hebert, 1983; Grimson and Lozano-Perez, 1984), Bolles and Horaud took a depth-first tree search as the basic matching strategy. Their scheme uses a minimum number of features to key into the data, first generating hypotheses and then taking care to ensure verification. [Note that Bolles and Cain (1982) had earlier used this technique in a 2-D part location problem.] This contrasts with much work (especially that based on the HT) which makes hypotheses but does not check them. (Note that forming the initial hypotheses is the difficult and computation intensive part of the work: researchers will therefore write about this aspect of their work and perhaps not state the minor amount of computation that went into confirming that objects had indeed been located: note also that in much 2-D work, images can be significantly simpler and the size of the peak in parameter space can be so large as to make it virtually certain that an object has been located—thus rendering verification unnecessary.)

16.12 CONCLUDING REMARKS

To the layman, 3-D vision is an obvious and automatic result of the fact that the human visual system is binocular and presumes both that binocular vision is the only way to arrive at depth maps and that once they have been obtained the subsequent recognition process is trivial. However, what this chapter has actually demonstrated is that neither of these commonly held views is valid. First, there are a good many ways of arriving at depth maps, and some of them are available using monocular vision. Second, the complexity of the mathematical calculations involved in locating objects and the amount of abstract reasoning involved in obtaining robust solutions—plus the need to ensure that the latter are not ambiguous—are taxing even in simple cases, including those where the objects have well-defined salient features.

Despite the diversity of methods covered in this chapter, there are certain important themes: the use of "trigger" features, the value of combining features into groups that are analyzed together, the need for working hypotheses to be generated at an early stage, the use of depth-first heuristic search (combined where appropriate with more rigorous breadth-first evaluation of the possible

interpretations), and the detailed verification of hypotheses. All these can be taken as parts of current methodology; *details*, however, vary with the data set. More specifically, if a new type of industrial part is to be considered, some study must be made of its most salient features: then this causes not only the feature detection scheme to vary but also the heuristics of the search employed—and also the mathematics of the hypothesis mechanism. The reader is referred to the following chapter for further discussion of object recognition under perspective projection.

Although the previous two sections have concentrated on object recognition and have perhaps tended to eschew the value of range measurements and depth maps, it is possible that this might give a misleading impression of the situation. In fact, there are many situations where recognition is largely irrelevant but where it is mandatory to map out 3-D surfaces in great detail. Turbine blades, automobile body parts or even food products such as fruit may need to be measured accurately in 3-D: in such cases it is known in advance what object is in what position, but some inspection or measurement function has to be carried out and a diagnosis made. In such instances, the methods of structured lighting, stereopsis or photometric stereo come into their own and are highly effective methods. Ultimately, also, one might expect that a robot vision system will have to use all the tricks of the human visual system if it is to be as adaptable and useful when operating in an unconstrained environment rather than at a particular worktable.

This has been a preliminary chapter on 3-D vision, setting the scene for Parts 4 and 5. In particular, Chapter 17, Tackling the Perspective n-Point Problem will be devoted to a careful analysis of the distinction between weak and full perspective projection and how this affects the object recognition process; Chapter 18, Invariants and Perspective will aim to show something of the elegance and value of invariants in providing short cuts around some of the complexities of full perspective projection; Chapter 19, Image Transformations and Camera Calibration will consider camera calibration and will also consider how recent research on inter-relating multiple views of a scene has allowed some of the tedium of camera calibration to be by-passed; and Chapter 20, Motion will introduce the topic of motion analysis in 3-D scenes.

Conventional wisdom indicates that binocular vision is the key to understanding the 3-D world. This chapter has shown that the correspondence problem makes the practice of binocular vision tedious, although the solutions it provides are only depth maps and require further intricate analysis before the 3-D world can fully be understood.

16.13 BIBLIOGRAPHICAL AND HISTORICAL NOTES

As noted earlier in the chapter, the most obvious approach to 3-D perception is to employ a binocular camera system. Burr and Chien (1977) and Arnold (1978) showed how a correspondence could be set up between the two input images by

use of edges and edge segments. Forming a correspondence can involve considerable computation: Barnea and Silverman (1972) showed how this problem could be alleviated by passing quickly over unfavorable matches. Likewise, Moravec (1980) devised a coarse-to-fine matching procedure which arrives systematically at an accurate correspondence between images. Marr and Poggio (1979) formulated two constraints—those of uniqueness and continuity—that have to be satisfied in choosing global correspondences: these constraints are important in leading to the simplest available surface interpretation. Ito and Ishii (1986) found that there is something to be gained from three-view stereo in offsetting ambiguity and the effects of occlusions.

The structured lighting approach to 3-D vision was introduced independently by Shirai (1972) and Agin and Binford (1973, 1976), in the form of a single plane of light, whereas Will and Pennington (1971) developed the grid coding technique. Nitzan et al. (1977) employed an alternative light detecting and ranging scheme for mapping objects in 3-D; here short light pulses were timed as they traveled to the object surface and back.

Meanwhile, other workers were attempting monocular approaches to 3-D vision. Some basic ideas underlying shape-from-shading date from as long ago as 1929, with Fesenkov's investigations of the lunar surface: see also van Digellen (1951). However, the first shape-from-shading problem to be solved both theoretically and in an operating algorithm appears to have been that of Rindfleisch (1966), also relating to lunar landscapes. Thereafter, Horn systematically tackled the problem both theoretically and with computer investigations, starting with a notable review (1975) and resulting in prominent papers (e.g., Horn, 1977; Ikeuchi and Horn, 1981; Horn and Brooks, 1986), an important book (Horn, 1986) and an edited work (Horn and Brooks, 1989). Interesting papers by other workers in this area include Blake et al. (1985), Bruckstein (1988) and Ferrie and Levine (1989). Woodham (1978, 1980, 1981) must be credited with the photometric stereo idea. Finally, the vital contributions made by workers on computer graphics in this area must not be forgotten—see, for example, Phong (1975), Cook and Torrance (1982).

The concept of shape-from-texture arose from the work of Gibson (1950) and was developed by Bajcsy and Liebermann (1976), Stevens (1980), and notably by Kender (1980), who carefully explored the underlying theoretical constraints.

The paper by Barrow and Tenenbaum (1981) provides a very readable review of much of this earlier work. 1980 marked a turning point, when the emphasis in 3-D vision shifted from mapping out surfaces to interpreting images as sets of 3-D objects. Possibly, this segmentation task could not be tackled earlier because basic tools such as the HT were not sufficiently well developed. The work of Koenderink and van Doorn (1979) and Chakravarty and Freeman (1982) was probably also crucial in providing a framework for interpretation schemes to be developed by using potential 3-D views of objects. The work of Ballard and Sabbah (1983) provided an early breakthrough in segmentation of real objects in 3-D and this was followed by vital further work by Faugeras and Hebert (1983),

Silberberg et al. (1984), Bolles and Horaud (1986), Horaud (1987), Pollard et al. (1987), and many others.

Other interesting work includes that of Horaud et al. (1989) on solving the perspective 4-point problem (finding the position and orientation of the camera relative to known points): for further references on this topic, see Section 17.6.

Though already a well worked-through topic, research on finding vanishing points proceeded further in the 1990s (e.g., Lutton et al., 1994; Straforini et al., 1993; Shufelt, 1999). Similarly, stereo correlation matching techniques were still under development, to maintain robustness in real-time applications (Lane et al., 1994).

Since 2000, work on stereo vision has continued unabated as a main-line topic (e.g., Lee et al., 2002; Brown et al., 2003), but Horn's approach to shape from shading has been largely superseded. One new technique is the Green's function approach to shape from shading (Torreão, 2001, 2003), whereas local shape from shading has been used to improve the photometric stereo technique (Sakarya and Erkmen, 2003). Photometric stereo has itself been developed considerably further in a new 4-source technique capable of coping with highlights and shadows (Barsky and Petrou, 2003). Another development is the application of shape from shading to radar data—a translation that required significant new theory (Frankot and Chellappa, 1990; Bors et al., 2003). Finally, a thoroughgoing new approach to the whole study of 3-D vision and its dependence on the light field has been initiated (Baker et al., 2003). This paper starts by comparing what can be learned from (1) stereo vision and (2) a shape from silhouette approach (observing object silhouettes from all directions in the given light field). An important conclusion is that the shapes of Lambertian objects can be uniquely determined with n-camera stereo, unless there are regions of constant intensity present: indeed, constant intensity is found *always* to lead to ambiguity. Essentially, this is because there may be a concavity whose light properties outside the concavity hull will be indistinguishable from those of the hull itself (Laurentini, 1994). Finally, we note that the paper by Baker et al. (2003) is important not only in giving a fresh view of the problems of 3-D vision in general, and shape from shading in particular, but also in demonstrating certain open questions.

16.13.1 MORE RECENT DEVELOPMENTS

Although the complexity of the image acquisition needed for photometric stereo should perhaps have made it relevant only during the early stages of the subject, the opposite now seems to be the case. First, Hernandez et al. (2011) indicate why this could be so: if a set of lights of different colors is arranged, there is no need to switch them, as the different color channels can be handled independently. However, this means that normally only three lights can be used, so the Barsky and Petrou (2003) 4-light technique cannot be employed, and this makes it difficult to confirm the interpretations obtained using the (usual) minimum number of three lights—an especially important factor when shadows occur.

Nevertheless, Hernandez et al. (2011) are able to use regularization methods that cope with as few as two light sources. Wu and Tang (2010) employ the opposite approach of using a dense image set and exploit the resulting data redundancy to determine how well the observations fit a Lambertian model. An expectation maximization approach is used to interpret the data in two stages, concentrating first on surface normals and then on surface properties including orientation discontinuities. The approach is robust and produces good reconstruction results. Goldman et al. (2010) note that most objects are composed of only a small number of fundamental materials: they therefore constrain pixel representations to at most two such materials, and thereby recover not only the shape but also material bidirectional reflectance distribution functions and weight maps. McGunnigle and Dong (2011) propose a photometric stereo method in which a conventional four-light scheme is augmented with coaxial illumination. Their investigations show that coaxial illumination makes photometric stereo more robust to shadow and specularity.

Chen et al. (2011) devise a fast stereo matching algorithm that uses a global graph-cuts framework, but which is as efficient as some local approaches. By concentrating on region boundaries and cleverly limiting the number of disparity candidates, the number of vertices in the constructed graph is significantly reduced. As a result, promising disparities can readily be selected and partial occlusions can be handled efficiently, thereby improving stereo matching speed.

16.14 PROBLEMS

1. Prove that all epipolar lines in one image plane pass through the point that is the image of the projection point of the alternate image plane.
2. What is the physical significance of the straight line contour in gradient space (see Fig. 16.9B)?
3. Sketch a curve of the function $\cos^m\theta$. Estimate what value m would have to have for 90% of the R_1 component to be reflected within 10° of the direction for pure specular reflection.
4. An alien has three eyes. Does this permit it to perceive or estimate depth more accurately than a human? What would be the best placement for a third eye?
5. A cube is viewed in orthographic projection. Show that although the cube is opaque, it is easy to compute the theoretical position of its centroid in the image. Show also that the orientation of the cube can be deduced by considering the apparent areas of its faces. If the contrast between the faces becomes so low that only a hexagonal outline is seen, show that ambiguities will arise in our knowledge of the orientation of the cube. Are ambiguities specific to cubes, or do they arise with other shapes? Why?

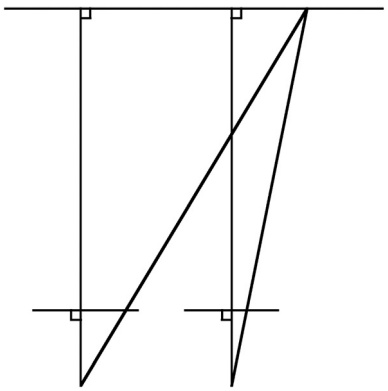

FIGURE 16.P.1

Geometry of a binocular imaging system

6. a. A feature at (X, Y, Z) appears at locations (x_1,y_1), and (x_2,y_2) in the two images of a binocular imaging system. The image planes of both cameras lie in the same plane, f is the focal length of both camera lenses, and b is the separation of the optical axes of the lenses. Label Fig. 16.P.1 appropriately: by considering pairs of similar triangles, show that:

$$\frac{Z}{f} = \frac{X + b/2}{x_1} = \frac{X - b/2}{x_2}$$

b. Hence, derive a formula which can be used to determine depth Z from the observed disparity.

7. Give a full proof that the error with which the fractional depth Z in a scene can be computed is (1) proportional to pixel size, (2) proportional to Z, and (3) inversely proportional to the baseline b between the stereo cameras. What other parameter appears in the final formula? Determine under what pair of conditions two very tiny cameras fabricated by nanotechnological methods could still perform viable depth measurement.

8. a. Draw a diagram which shows that the ordering of visible points is normally the same in both images seen by a binocular vision system.

b. An object has a semitransparent front surface through which an interior feature F is just visible. Show that the ordering of the features in the two views of the object may be sufficient to prove that F is inside, or perhaps behind, the object.

9. a. State the conditions under which matt surfaces may properly be described as "Lambertian." Show that the normal at a point on a Lambertian surface must lie on a cone of directions whose axis points to the point course of illumination. Show that a minimum of three independent light sources will be needed to identify the exact orientation of a matt surface. Why might four light sources help to determine surface orientation for a surface of unknown or nonideal properties?

b. Compare the effectiveness of binocular vision and photometric stereo if it is desired to obtain a depth map for each object in a scene. In each case, consider the properties of the object surface and the distance from the observer.

10. a. Compare the properties of matt surfaces with those which exhibit "normal" specular reflection. Matt surfaces are sometimes described as "Lambertian." Describe how the brightness of the surface varies according to the Lambertian model.

b. Show that for a given surface brightness, the orientation of any point on a Lambertian surface must lie on a certain cone of orientations.

c. Three images of a surface are obtained on illuminating it in sequence by three independent point light sources. Show with the aid of a diagram how this can lead to unambiguous estimates of surface orientation. Would surface orientation of any points on the surface *not* be estimated by this method? Are there any constraints on the allowable positions of the three light sources? Would it help if *four* independent point light sources were used instead of three?

d. Discuss whether the surface map that is obtained by shape from shading is identical to that obtained by stereo (binocular) vision. Are the two approaches best applied in the same or different applications? To what extent is the application of structured light able to give better or more accurate information than these basic approaches?

e. Consider what further processing is required before 3-D objects can be recognized by any of these approaches.