9

DEPTH MEASUREMENT AND ANALYSIS

Many scene analysis difficulties are caused by the loss of direct depth information in a single picture. Segmentation techniques are complicated because depth discontinuities are being implicitly inferred from intensity, color, or texture discontinuities. Shape analysis is made difficult by the changes in the projected shape caused by perspective.

Depth of visible surfaces can be measured by passive use of multiple views, as in stereo vision, or by active control of the illumination of an object. Humans are able to perceive relative depth from a single view, as in perceiving a photograph. Such inference is believed to use many complex "cues" such as occlusion, shadows, texture gradients, and size of familiar objects. Object, or viewer, motion provide another important cue for object detection and segmentation.

Only the depth of the visible surfaces can be inferred from single views or narrowly spaced stereo views. Such data is sometimes called a 2 1/2-D model of the scene. A number of views are required to obtain complete 3-D models of all surfaces. In our normal perception such complete models are, of course, not available.

9.1 STEREO AND MOTION

Each point of an opaque object's image corresponds to one point on the object. This object point must be along the ray joining the image point and the focal point for an ideal lens imaging device, but the distance to the object along this ray is unknown. However, if the object is viewed from another perspective and the same point is visible in both views, then it must be along the intersection of the rays determined from the two views, and its three-dimensional position can be computed. For example, in Fig. 9-1, the points P and Q are constrained to be along rays C_1P_1 and C_1Q_1, respectively, from the information available in the left view, and along C_2P_2 and C_2Q_2 by the second view. The two constraints together determine the positions of P and Q.

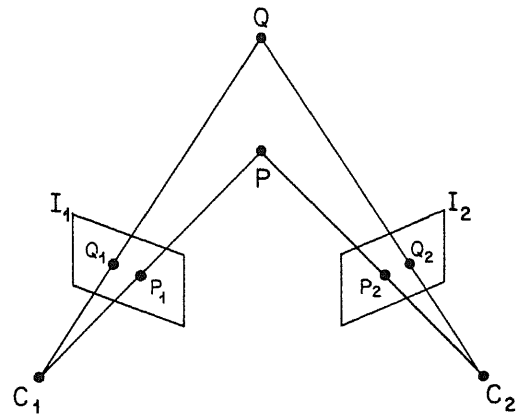

Figure 9-1: Determining position by stereo triangulation

Stereo depth measurement requires the following two operations.

1. Determination of the point pairs in the two images that correspond to the images of the same object point. This is known as the *correspondence* problem. The translation between corresponding points in two images is known as the *disparity* of the point pair. Note that the image of a small surface on the object may be different in the two images because of the changes in perspective and in surface reflectivity with the imaging angles. Moreover, some of the points in one image may not be visible in the other.

2. Determintion of the three-dimensional object point from the two image points by *triangulation.* This operation requires the relative positions and orientations of the cameras for the two views.

The triangulation problem is simple if the two cameras are calibrated with respect to each other. If enough corresponding points are obtained, the relative camera parameters can be determined by minimizing the shortest distances between the corresponding lines that should ideally intersect. The algebraic details are complex, but they can be worked out following the procedures of Chapter 3. Details may be found in [1, 2].

Correspondence between two views can be established by matching specific features, such as corners, in two views or by matching small regions by correlation without identifying any features. Matching of specific features is potentially more efficient, as their number is small. The features to be used should be distinct and also invariant to small changes in the viewing angle. Note that line segments are distinct in orientation only, and not in position. Corners are distinct in both of these parameters.

Use of features for stereo correspondence has been limited, owing partly to the difficulty of extracting reliable features. Burr and Chien [3], Arnold [4], and Baker and Binford [5] have used edges and edge segments and their associated properties such as length and contrast. Marr and Poggio [6] and Grimson [7] have described the use of "zero-crossings," of an image convolved with a Laplacian-Gaussian mask (as in Section 7.1.3). Ganapathy [8] and Underwood and Coates [9] have described matching of features using invariant properties of polyhedral scenes; Shapira and Freeman describe a technique for matching of imperfect, curved object scenes [10].

Searching for correspondence by similarity of regions in two images does not require any previous analysis of the images, but the search is more expensive. Two measures of similarity between regions are commonly used: the cross-correlation, as defined in Eq. (2-2), and the sum of the squares of the differences. The cross-correlation coefficient is independent of the contrast but requires more computation. Sum of the squares of the differences is adequate if the two views are taken under similar viewing conditions.

For a region centered around a given point in an image, the second image can be searched for a window that is most similar to the first based on one of the above measures. This operation needs to be performed for windows around each point in the first image whose depth is to be measured. Note that this is possible only for those windows that contain sufficient information for matching. Such windows should be nonhomogeneous and contain unique features. The search for a match can be quite expensive, if conducted exhaustively. Different ways to reduce this search are described in Section 9.1.1.

The correspondences found by measurement of local similarities, by region correlation, or by feature matching may occasionally be ambiguous or in error. Some of these ambiguities can be resolved by examining the disparities of several points in context, as described in Section 9.1.2.

Detection of motion has similarities with the stereo correspondence problem. However, 3-D data is more difficult to obtain from object motion. A brief survey of motion detection and analysis techniques is given in Section 9.1.3.

9.1.1 Correspondence Search

As the disparity in two images depends on the distance of the objects from the camera, if the limits of the range to the object are known, search for disparities can be limited correspondingly. More restrictions can be derived by observing that a point in one image constrains the corresponding object point to lie along a certain straight line in 3-D space and that the image of this line in a second view is also a line in that view. Thus, the search for correspondence with the image of a point need only be along a line. To account for errors in the knowledge of the relative orientations and positions of the two cameras, the search needs to be conducted in a narrow band. Again, the search distance along the line can be limited if the limits on the range to the object from the camera are known.

Barnea and Silverman used an interesting variation by not computing a similarity measure for the entire window but instead basing the decision on the rate at which the differences in the two windows accumulate [11]. A match with high error when only a few pixels have been examined is abandoned in favor of alternative matches.

Local context can be used to speed up the search, assuming that the surfaces in 3-D are smooth except at a small number of discontinuities. Disparity of the neighboring points can thus be expected to be similar, and the search can use the neighbor's disparity for a starting point.

Two search procedures using additional simplifications are described below.

Use of multiple views. The accuracy of depth measurement by stereo depends directly on the angle, hence the disparities between the two views. However, a large angle and larger disparities require a search over larger image areas. Use of multiple views ~~between two stereo views~~ allows high accuracy without increase in search time.

Availability of multiple views offers several alternatives for searching for correspondences. Assume that the search for disparities between the two extreme views of a series of k views can be limited to a band n pixels long and m pixels wide, as discussed above. The search

for correspondences between two adjacent views can then be limited to a band n' pixels long and m' pixels wide, where $n' = (n/k-1)$ and $m' = (m/k-1)$. Let the successive views be called $view_1$, $view_2$, ..., $view_k$.

Disparities between extreme views can be determined by chaining through the intermediate views. Consider the matching of a particular, chosen region in $view_1$. Assume that a match has been found for this region between $view_1$ and $view_2$. To determine the best match for the same region in $view_i$ with a region in $view_{i+1}$, the search need be conducted only in an n'-by-m' band, centered at the center of the region of best match in the previous view ($view_i$). This process begins with matching of all regions of interest in $view_1$ with regions in $view_2$ and continues until the last view has been considered. An alternative is to match views i and $i+1$ at each step, where $1 < i \lessdot k-1$, and sum the disparities. For the former method, however, only crude disparities need to be found at all but the last step, and the errors of matching do not accumulate.

Assuming exhaustive searches in the specified bands, the total number of matches examined for all steps will be $(k-1)(n'm')$ which is equal to $(nm)/(k-1)$. This is a saving of a factor of $(1/k-1)$ compared to the direct search between the extreme views. If the width of the search band is assumed independent of the angle, then no savings in search results. However, use of multiple views may still be more reliable, as at each step the images differ by smaller amounts, and the search is constrained to smaller areas.

Use of multiple views, taken every 1/2 degree apart, for a total stereo angle of five to ten degrees, is described in [12]. Baumgart has used multiple views around an object to generate complete 3-D models for computer graphics applications [13].

Coarse to fine matching. A considerable saving in search time can be obtained by matching reduced-resolution images of the two views and refining the match in successively higher-resolution images. At each step of the matching, the search window can be maintained to be of the same size.

Such coarse-to-fine matching has been used by Moravec in a stereo program [14] and is similar to the use of reduced-resolution images for other purposes, as in planning procedures for edge detection and region segmentation. A possible difficulty of this approach is in loss of information in the reduced image and hence missing of some correspondences.

9.1.2 Global Correspondences

The correspondences computed by the local similarities may be ambiguous if more than one region in the same image has similar properties. Consider the left and right images in Fig. 9-2, consisting of three dark squares each, as marked. Each square in one image is similar to any of the three in the other. If we correspond L_1 and R_1, L_2 and R_2, and L_3 and R_3, the three squares will be computed to be at the same height above the background, shown by the filled circles in Fig. 9-2. If L_1 is matched with R_2, L_2 with R_3, and L_3 with R_1, then the computed heights of the squares will be as indicated by the cross marks in Fig. 9-2. Another possible interpretation is shown by the unfilled circles.

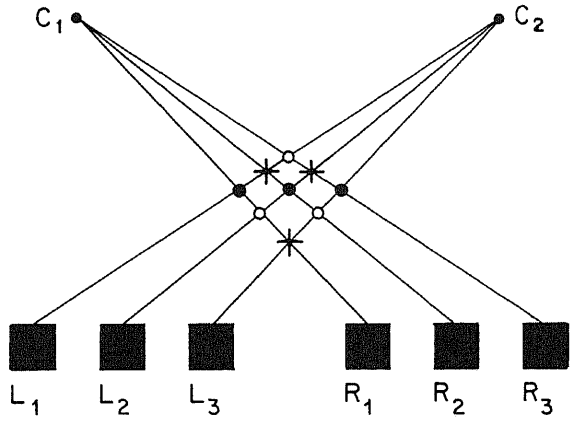

Figure 9-2: An example of ambiguous stereo correspondence

In spite of local ambiguities, our perception of stereo images is normally unambiguous. Apparently, certain correspondences are preferred as giving a total perceived surface with preferred properties. Julesz has suggested a "cooperative model" for stereopsis, where the final disparity values of each cell are influenced by disparities of neighboring pixels [15]. A dipole is associated with each pixel, pointing in the direction of the disparity. However, the direction of each dipole is also determined by its coupling by springs to neighboring dipoles. This model appears to have desirable qualitative properties but is incompletely specified and untested.

Marr and Poggio have suggested that two constraints should be satisfied in choosing global correspondences [6]:

1. *Uniqueness.* Each point in an image must be assigned at most one disparity value, as the corresponding object point has a unique

physical location. This constraint is not valid for transparent or translucent objects.

2. *Continuity.* Disparity values change smoothly, except at a limited number of depth discontinuities (since matter is cohesive).

The above two constraints lead to the choice of the simplest (most continuous) surface interpretation among the many alternatives. Marr and Poggio have described two techniques to implement these constraints.

Their first method is described by considering a three-dimensional collection of binary-valued cells, each plane corresponding to pixels with the same disparity. Let $C^t_{x,\,y,\,d}$ denote the state of a cell in this network, and its being *on* corresponds to the pixel (x, y) in the left image having a disparity d (for convenience, all disparities are considered to be in the x direction). Cells in the network either reinforce or inhibit the response of the neighboring cells. Marr and Poggio suggested the following iterative algorithm:

$$C^{t+1}_{x,y,d} = \sigma\left[\sum_{\substack{x',y',d \\ \in S(x,y,d)}} C^t - \varepsilon \sum_{\substack{x',y',d \\ \in O(x,y,d)}} C^t_{x',y',d} + C^0_{x,y,d}\right] \qquad (9\text{-}1)$$

where superscript t indicates time or iteration, $S(x, y, d)$ is the excitatory neighborhood and $O(x, y, d)$ the inhibitory neighborhood of cell (x, y, d), ϵ is a constant parameter, and σ is a threshold function. $S(x, y, d)$ is taken to be a circular neighborhood of a certain size, (such as with a radius of five pixels) of (x, y, d) in the same disparity layer in the network, indicating a preference for the continuity of the disparity values. Cells with the same (x, y) values but with different disparities comprise the inhibitory neighborhood corresponding to the uniqueness criterion. These neighborhoods correspond to the lines of sight from the two views, if the network cells are actually located at the 3-D positions indicated by the image (x, y) coordinates and disparity d.

Marr and Poggio have described experiments with random dot stereograms and claim convergence for a wide range of parameter values in Eq. (9-1). However, these examples are characterized by relatively uniform disparities and large step discontinuities, and the performance on real images remains unclear. A similar iterative algorithm was described earlier by Dev [16].

The second algorithm of Marr and Poggio is more heuristic in nature, and a specific implementation is described by Grimson [7]. Matching between two images is between zero-crossings obtained by

Laplacian-Gaussian edge-detecting masks (as described in Chapter 7). A match for a zero-crossing in one image is centered in three locations (or three "pools") in the other image, one corresponding to the expected disparity, and the others on two sides of it. If two matches are found around the same location (or in the same pool), it is decided that no match exists. If a match is found in more than one pool, then the pool that gives a disparity value similar to disparities of other points in the neighborhood of this point is chosen. (Arnold describes a similar approach, requiring edge elements linked to a central edge to have disparities nearly equal to that of the central element [4]. Baker and Binford give additional contraints based on the connectivity of the edges [5].)

The expected disparity needed in the procedure above is obtained successively by starting from larger masks and narrowing to the smaller masks. Initially, for the largest mask, the expected disparity is assumed to be zero.

The most common use of stereo is perhaps in mapping of terrain elevation contours from aerial images. In stereo matching over natural textured areas and smoothly varying terrains, ambiguities in local correspondences are few and can be resolved by the continuity of the disparity values. Stereo correspondence is more difficult in the presence of sharp changes in range as for steep terrain or in the presence of cultural features such as buildings, and for indoor scenes with occluding objects. In such cases, no correspondences may exist for points at the boundaries or at surfaces of steep slopes. The above-described global correspondence techniques have been tested on a variety of indoor and outdoor scenes, but only those of limited complexity in the number of occluding objects. Note that for nontextured objects, stereo range information is available only at particular distinct points, such as at corners or edges.

9.1.3 Motion Detection and Analysis

Motion of an object in a dynamic scene is a very strong cue to its presence, and some animals, such as frogs, are thought to rely solely on motion in search of their prey. Motion of an object can be detected by finding corresponding points in a sequence of images. However, the same constraints cannot be used for limiting the search or establishing the global correspondences. If the object motion is small and much of the scene is static, the moving points can be detected simply by differencing the two views [17]. A spatial cluster of points with high differences may be used to distinguish the moving points from the points whose intensity changes due to noise.

Detection of motion by first segmenting objects and then matching them in a sequence of views is simple in principle, but difficulties arise owing to errors of segmentation and to possible rotation or occlusion of objects. A complete survey of such and other motion-detection techniques may be found in [18].

Some techniques have tried to use the rate of change of intensity at the pixels in an image for the estimation of the velocity of moving objects and for their segmentation from the background. Note that for objects with homogeneous surfaces, only the points at the boundaries with a component normal to the object motion have a nonzero intensity change.

For a given point, an incremental change in intensity di, due to a spatial shift $d\mathbf{s}$, is given by

$$di = -\mathbf{G} \cdot d\mathbf{s} \tag{9-2}$$

where $\mathbf{G}$ is the spatial intensity gradient and the negative sign is due to the motion of the underlying surface (rather than the observer). Taking derivatives with respect to time, we get

$$\frac{di}{dt} = -\mathbf{G} \cdot \mathbf{V} = -(G_x V_x + G_y V_y) \tag{9-3}$$

where $\mathbf{V}$ is the object velocity with components V_x and V_y along the x and y axes, and G_x and G_y are the components of the gradient $\mathbf{G}$.

Thus, given di/dt and $\mathbf{G}$, constraints are placed on the object velocity, but the components V_x and V_y cannot be inferred directly. Fennema and Thompson used clustering of points in the (V_x, V_y) space to infer the magnitude and the direction of the velocity [19]. Thompson also combined such velocity information with intensity information using a region-growing approach to segment objects [20]. Another important technique for segmenting objects, given points of large intensity change, is described in [21] and is based on an analysis of expected intensity changes when two surfaces move relative to each other.

Ullman has implemented a system of correspondence for motion by matching of local edges and line segments [22]. As in stereo, the local matches may be ambiguous. Ambiguities are resolved by testing whether a given set of matches could correspond to a *rigid-body* motion. Ullman shows that given three distinct orthographic views of four noncoplanar points in a rigid configuration, the structure and motion compatible with the three views are uniquely determined. Inference of nonrigid motion from moving light displays is described in [23].

Apparent velocities of the points in an image, called *optical flow*, can also be useful for inferring the three-dimensional structure of the object surfaces under certain simplifying assumptions. These techniques have been tested only on very limited, simple scenes and will not be described here; the reader is referred to [24] and [25], which also contain extensive bibliographies.

9.2 ACTIVE RANGING

Range measurement by stereo requires only *passive* observation of the pointscene from multiple viewpoints. Range can be measured more easily if active control of the illumination is allowed. Two approaches are described below: triangulation of a pattern of light and measurement of time of flight of a signal to the object.

9.2.1 Triangulation Ranging

Consider viewing an object illuminated by a single ray of light of known 3-D position and orientation. The illuminated point on the object will form a single-point image on the camera screen. Given this point in the image, we can constrain the illuminated object point to be on a line through the image point and the camera center, computed from a known camera transform. As the object point must also lie on the illuminating ray, its position can be determined by the intersection of the two constraining lines. This is known as triangulation ranging, and is similar to the stereo ranging shown in Fig. 9-1, with one of the cameras replaced by a collimated source of light. To obtain positions of all points on the object, the object needs to be scanned by varying the position and the orientation of the illuminating beam.

The scanning of an object can be speeded up if it is illuminated by a plane of light rather than a single collimated beam. The object is then illuminated along a planar curve in 3-D and forms a line image on the imaging plane as shown in Fig. 9-3. For each point along this image curve, a straight line to the corresponding object point can be associated as before. The object must lie along the intersection of straight line and the known illuminating plane. To obtain other object points, the object can be scanned by a series of parallel planes.

This procedure will give poor resolution for object surfaces nearly parallel to the illuminating plane. Two mutually orthogonal series of scans may be used to overcome this difficulty. Agin and Binford [26] and Shirai [27] first described such ranging implementations independently. The plane of light is produced by illuminating a linear

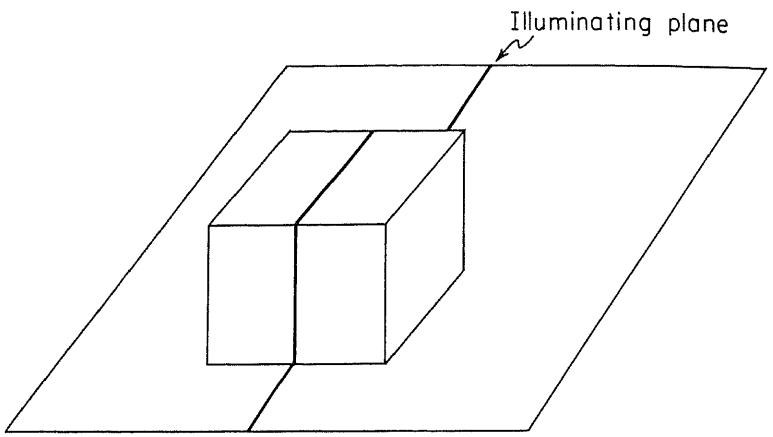

Figure 9-3: Illumination of an object by a plane of light

slit or by passing a collimated beam through a cylindrical lens. In the latter case, the desired orientations can be obtained by rotating the lens. A series of parallel scans can be obtained by a scanning mirror.

Figure 9-4 shows a TV camera image of a scene containing two objects. Figure 9-5 shows laser scans in two directions. Boundaries of objects can be derived by linking the end points of these scans, as they correspond to the depth discontinuities in the scene. A boundary may not cross any of the scans or itself. Figure 9-6 shows boundaries derived from Fig. 9-5 (from [28]). Note that the objects are well separated from the background, but the touching parts of the doll and the snake are not.

The ranging process can be further speeded by illuminating the objects with a grid of a predetermined pattern, say a light square grid on a dark background. If such a grid were to illuminate polyhedral objects, the images of parallel lines on a face would be parallel, ignoring perspective distortions, but the angles of these lines would depend on the orientation of the surface. Thus, discontinuities in surface depth and slope would result in discontinuities in the image lines and their slopes, respectively. Will and Pennington have described a *grid coding* technique that uses a square grid illumination, and different faces of polyhedral objects are separated by providing separate peaks in a Fourier transform of the image [29]. They suggest that other grid patterns may be suited for other specific objects. Unfortunately, precise 3-D positions of object points cannot be determined by grid illumination, unless the individual grid lines can be identified, say, by color.

Figure 9-4: A TV picture of a scene

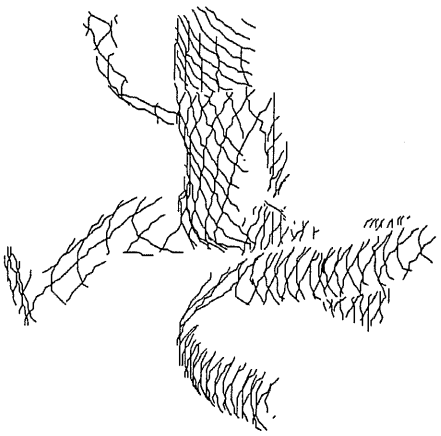

Figure 9-5: Laser scans for the scene of Fig. 9-4

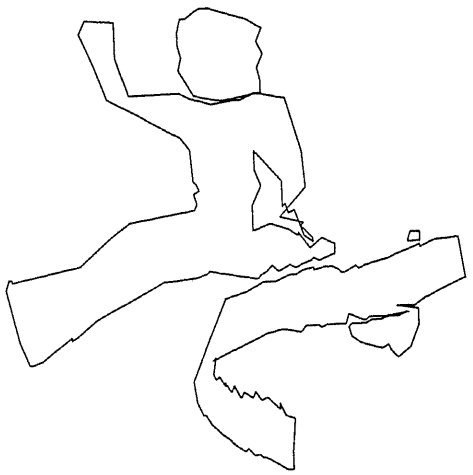

Figure 9-6: Boundary derived from the scans of Fig. 9-5

9.2.2 LIDAR Ranging

Distance of an object from the observer can be determined by transmitting an electromagnetic signal and measuring the time interval for the reflected signal to come back, as in the use of RADAR for aircraft range measurement. For objects only a few meters away, signals need to be of shorter wavelength than are commonly used in RADAR. The visible light spectrum is suitable, and such systems have been called LIDAR (Light Detecting and Ranging) systems.

Range can be measured by transmitting a short pulse of light and measuring the time interval for the return path, or by transmitting a continuous wave signal and measuring the phase shift. A pulsed system is described in [30] and a continuous wave system in [31]. The latter system is shown schematically in Fig. 9-7. Discussion of details of electronics for time or phase measurements is beyond the scope of this book.

9.3 SEGMENTATION USING RANGE

Use of range data enables us to segment objects in a scene by the property that different objects are separated in 3-D space, rather than by discontinuities in their surface properties. Of course, objects that touch each other or rest on top of another cannot be segmented so easily. Segmentation using range data basically requires isolating the discontinuities in range. Analogous to the processing of intensity data,

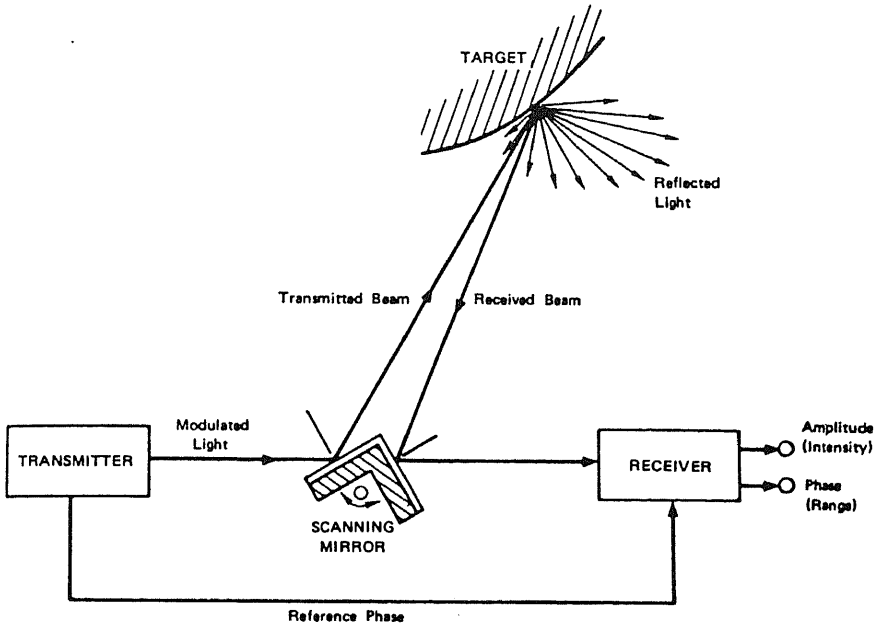

Figure 9-7: Block diagram of a laser ranging system (from Nitzan et al. [31]) (© 1977 IEEE)

either edge or region types of methods may be used.

9.3.1 Boundary Detection

Discontinuities in range can often be found simply by a large jump in the range values, as different objects are separated by distances that are large compared to the range resolution. Connecting such discontinuities gives the "jump boundaries." Certain boundaries, such as the intersection of two planar surfaces, will not be detected as jump boundaries. Use of first-derivative discontinuities, as is common for intensity edge detection, is also not useful, as any slanting surface has a high range derivative. Such boundaries can be detected by using the second derivative, but with increased probability of errors due to noise in the range data.

Boundaries obtained by detecting range edges can usually be expected to be more robust than those using the intensity data.

However, these boundaries may still not be complete and contain gaps and spurious edges. In some cases, these gaps can be filled by using the additional information available in range data (see [32]). Duda and Nitzan have also suggested the use of registered intensity data, as certain boundaries may be more apparent in intensity data [33]. Sugihara has used constraints on edges at vertices of polyhedral objects (given by the Huffman-Clowes theory described in Chapter 4), to hypothesize and verify the missing lines, and has also described extensions for curved objects [34].

9.3.2 Detection of Planes and Surfaces

Surfaces of known shape, such as planes, can be detected directly in the range data. However, detection of such surfaces is more complicated than testing if a given set of points lies on such a surface.

The easiest surfaces to detect are horizontal planes, since the only unknown is the height, z. As the points on the same horizontal surface should have the same height, a histogram of the z values can be expected to have peaks corresponding to such surfaces. Another method is to take slices of the points in various ranges of height and find the points belonging to a single surface (see [31, 33]).

Detection of vertical surfaces involves an extra degree of freedom. One approach is to project the points on a horizontal plane (x–y). The vertical planes should project into straight lines, which can be detected by a Hough transform or other methods described in Chapter 7.

Detection of planes of arbitrary orientation is more difficult. In principle, these planes can be detected by the use of a Hough transform, but the dimensionality of the transform space is likely to cause computational difficulties. Duda, Nitzan, and Barrett use peaks in intensity histograms, after the horizontal and vertical surfaces have been removed, to identify potential candidates for other planar surface, and they verify the hypothesized regions using range data [33].

Presence of planes could also be detected by computing normals to the local surfaces around each point and by clustering the orientations of these normals. This method may be sensitive to noise, as a derivative operation is required, and the directions may be in error near the boundary of two intersecting planes.

Curved surfaces of a given type could also be detected by a Hough transform approach, but the computational requirements are further increased because of the larger number of parameters. Oshima and Shirai have used an alternative approach [35]. They first group points into small surface elements and fit a plane to them. These elements are then merged into larger approximately planar regions, which are

classified into plane, curved, or undefined classes. The curved regions are further merged if the larger regions fit a quadratic surface. This approach is akin to the region-growing approach for intensity-image segmentation.

9.4 SHAPE FROM SHADING

Humans are able to extract depth information, at least in a relative and qualitative sense, from single images. Among the many cues believed to be used are occlusion, texture gradients, size of familiar objects, and smooth changes in shading. Our understanding of these processes is very preliminary, and most of the machine implementations apply in very restricted cases only. We first consider shape from shading.

In normal segmentation using edge or uniform region analysis, smooth variations in surface intensity are viewed as a source of difficulty. However, these variations can provide clues to the 3-D shape of the surface. To perform such computation, we need a better understanding of the image-formation process.

For most surfaces, the proportion of incident light reaching the viewer is a function of the surface orientation. For a single light source the ratio can be represented as a function $\Phi(i, e, g)$ of the three angles i, e, and g, known as the incident, emittance (or view), and phase angles, respectively (see Fig. 9-8). The incident angle is between the local surface normal and the incident ray, the emittance angle between the surface normal and the ray to the viewer, and the phase angle between the incident and the emitted ray.

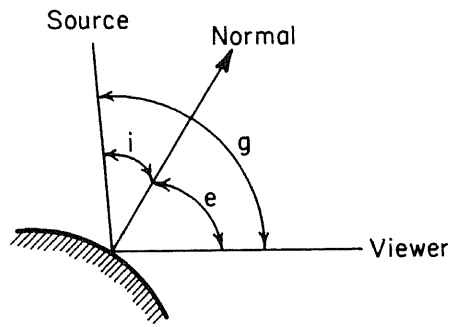

Figure 9-8: The incidence, emittance, and phase angles for determining reflectivity

For a given object point, illuminated by incident light intensity $A(\mathbf{r})$, the intensity of the corresponding image point $B(\mathbf{r}')$ is given by

$$B(\mathbf{r}') = A(\mathbf{r}) \cdot \Phi(i, e, g) \tag{9-4}$$

To compute shape from shading, we assume that $A(\mathbf{r})$ and $\Phi(i, e, g)$ are known and $B(\mathbf{r}')$ is obtained from the image. We wish to recover the surface orientation using (94). Horn has shown that this equation is a nonlinear partial differential equation, and has given a numerical technique for its solution under some simplifying conditions regarding the source and the camera locations [36]. The solution procedure starts from an initial contour and traces other contours satisfying the equation. The initial contours are derived from the brightest and the darkest points in the image. Horn was able to recover shape of smooth objects from a monocular view. (This technique is a generalization of an earlier technique to estimate shape of the lunar surface from a single view, where the reflectivity function has a simple, known form [37].)

Note that this technique requires complete knowledge of the light source and the reflectivity function $\Phi(i, e, g)$ for all combinations of the parameters. The reflectivity function may be known for some simple surfaces but must be measured carefully for others; mathematical models using surface properties are not very advanced. Also, in general, the surface materials of the of objects in a scene are not known a priori.

If an orthogonal, rather than a perspective, projection is used, a simpler and more elegant formulation using *reflectance maps*, also developed by Horn [38], is possible and is discussed below.

9.4.1 Reflectance Maps

Reflectance maps are based on the concept of gradient spaces, introduced in Chapter 4. The local surface normal of a surface $z = f(x, y)$, can be described by two parameters, $p = \partial z/\partial x$ and $q = \partial z/\partial y$. Parameters (p, q) constitute the gradient space. The shape-from-shading problem is essentially to estimate the (p, q) values for the points in an image.

In an orthographic projection, the viewing direction and hence the phase angle, g, is constant for all object points. Thus, for a fixed light source and viewing geometry, the reflected light and hence the image intensity depend only on the surface normal—that is, the gradient coordinates p and q. The reflectance map, $R(p, q)$ determines the image intensity as a function of p and q.

Let us consider a perfect Lambertian, or matte surface, which appears equally bright from any viewing angle. Here the reflectivity R is simply proportional to $\cos(i)$, where i is the incident angle. Also, if the source is near the viewer,

$$\cos(i) = \frac{1}{\sqrt{1 + p^2 + q^2}} \tag{9-5}$$

For this case, the reflectance map, plotted as a function of p and q, consists of concentric circles (see Fig. 9-9). A somewhat more complex relationship applies if the light source is positioned away from the viewer. For a source whose direction is given by the vector $(P_s, q_s, -1)$ it can be shown that

$$\cos(i) = \frac{1 + p_s p + q_s q}{\sqrt{1 + p^2 + q^2}\sqrt{1 + p_s^2 + q_s^2}} \tag{9-6}$$

The corresponding contours in the reflectance map for a particular source location are shown in Fig. 9-10. (The reflectance map is fixed for a given source position.)

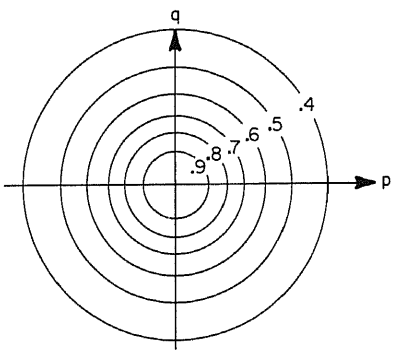

Figure 9-9: Contours in gradient space for a matte surface; source is at the viewer (adapted from Horn [38])

For a perfect mirrorlike reflector, the reflectance map is simply an impulse for one point in the (p, q) space. Reflectivity of most surfaces has both a mirrorlike specular component and a matte component.

The material in the maria (seas) of the moon also has a particularly simple reflectance function. Here, $\Phi(i, e, g) = \cos(i)/\cos$

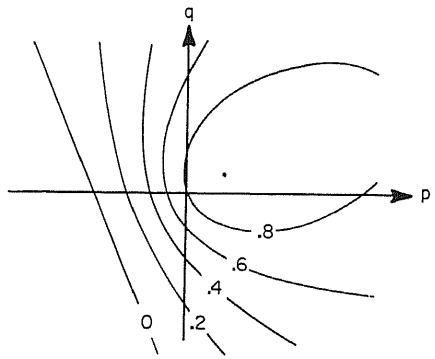

Figure 9-10: Another reflectance map for a matte surface; direction to the source is (0.7, 0.3) (adapted from Horn [38])

(*e*) and the corresponding reflectance map consists of parallel lines in the (*p, q*) plane.

The reflectance map representation is a convenient form for storing the reflectance properties and is useful for synthesizing images, given the surface of the shape. For the inverse process of shape from shading, a given intensity constrains the surface normal to be along a particular contour in the reflectance map. Assuming continuity of neighboring points on the surface, local shape can be determined as before. The basic process consists of tracing *base characteristics* in the image and computing corresponding *characteristics* in the reflectance map. The paths in the two spaces are related in the following manner. A step in the image space is perpendicular to the contour in the gradient space, and a step in the gradient space is perpendicular to the corresponding intensity contour in the image plane. (Further details are given in [38].) A parallel relaxation approach is described in [39].

Note that such processing requires an estimate of the initial conditions. It has been suggested that such an estimate may be derived from the surface boundaries. For smooth surfaces, at the external boundaries, the surface must be tangential, and the surface normal orthogonal to the line of sight [40]. Woodham gives some interesting constraints on the surface shape computation, if general properties of the surface, such as convexity, are known [41].

Woodham also suggests a novel approach to depth measurement, called *photometric stereo*, that uses two or more views of a scene by moving the light source rather than the camera [41]. For a given position of the light source, for each point in the image, a contour constraining the orientation of the surface at the point is given in the gradient space. Another position of the light source gives another such

contour, and the surface must be at one or more of the points of their intersection. More views can be used to improve accuracy and reduce any remaining ambiguities. Photometric stereo has not been tested extensively but is of potential use where illumination can be controlled, as for industrial applications. The advantage of photometric stereo over conventional stereo is that no correspondence solution is needed, as the images are in perfect registration. However, photometric properties of the surface—that is, its complete reflectance function must be known in advance and the surface must be uniform. Practical advantages of this scheme over the conventional stereo are unclear.

Another application of reflectance maps, described by Horn and Bachman, is for registration of satellite images taken with different sun angles [42]. One image is first converted to a reflectance or *albedo* image by conversion of image intensity into surface reflectance. This, however, requires knowledge of surface orientation, which is obtained by terrain data of the region, and of the reflectance function. A new synthetic image can now be obtained for any desired sun angle and used for registration.

The reflectance-map representation is also useful in explaining the occurrence of certain types of edges in polyhedral scenes, such as steps, roofs, and peak (as in Chapter 7). The peak edges are likely to be due to convex edges, the roof edges due to convex edges, and the step edges due to obscuring edges. Reflectance-map properties can also be used to extend the gradient space analysis of Mackworth (Section 4.3). (In the example of a trihedral vertex shown in Fig. 4-16, the size of the triangle in the dual space can be fixed by requiring the points A', B' and C' to lie on reflectance contours in the gradient space given by the intensities of the three faces A, B and C).

9.5 TEXTURE GRADIENTS

Smooth variations in texture, or *texture gradients*, can give clues to local surface shape. Figure 9-11 shows a picture of a brick wall. The bricks in the distant parts of the wall are smaller, in the image, and more closely spaced. The picture gives us a strong sense of the orientation of the wall, and the appropriate cue seems to be the gradient of the texture pattern. Figure 9-12 shows a synthetic image where a surface is suggested, again presumably by the changes in the elements of the perceived texture. The importance of texture gradients was suggested by Gibson as early as in 1950 [43].

Stevens has defined three causes of texture gradients [44]. The gradient may be caused by the variations in distance, called a *scaling gradient*, or by the variations in surface orientation, called a

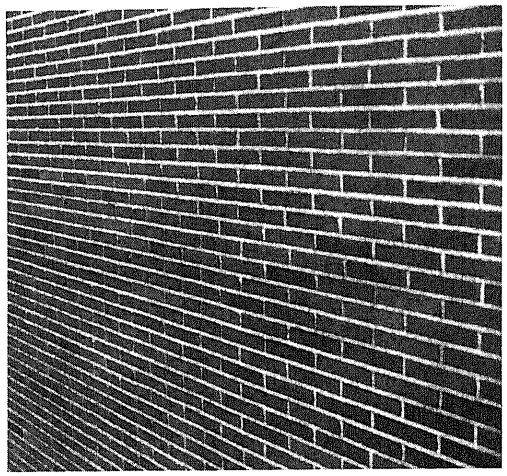

Figure 9-11: Texture gradient in a brick wall

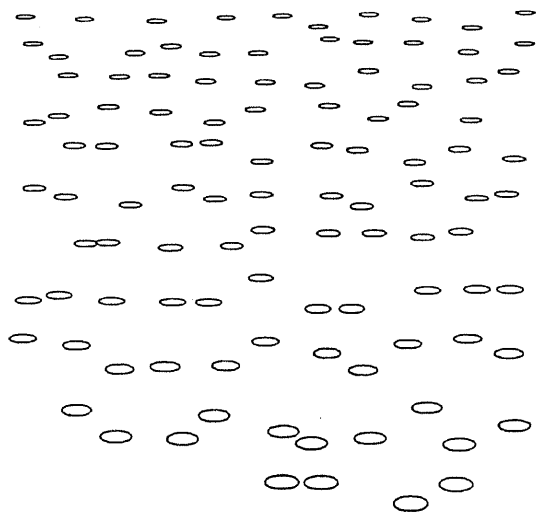

Figure 9-12: Perception of a surface from synthetic texture gradient (from Stevens [44])

foreshortening gradient, or by the variations in the physical texture itself. Normally, the physical texture is assumed to be constant.

Stevens has also suggested a modified gradient space representation for the normal to a surface that decomposes the effects of the scaling and the foreshortening gradients. He suggests the use of

polar coordinates τ and tan σ, of a point (p, q), in the gradient space as shown in Fig. 9-13. τ and σ are called the *tilt* and the *slant* of the surface. Tilt may be considered to specify "which way" the surface normal is oriented and slant to specify "how much." Consider a vertical image plane. A surface parallel to the image plane has zero tilt and slant. If the surface is rotated about a vertical axis, the tilt remains zero and the slant depends on the degree of rotation. Similarly, for all rotations about the horizontal axis, the tilt is 90 degrees and the slant gives the amount of rotation.

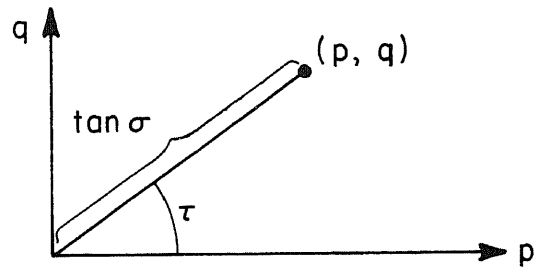

Figure 9-13: Tilt-slant representation of surface orientation

The scaling gradient of a texture can be computed from *characteristic dimensions* that are not foreshortened, and hence must be parallel to the image plane. In the image, these are in a direction normal to the direction of the most rapid change in any measure of texture regularity that is sensitive to scaling. This direction of most rapid change also gives the tilt angle τ. Alternately, the characteristic dimensions are locally aligned with the orientation of the greatest regularity. One or the other definition may be easier to apply for a particular texture. In Fig. 9-12 the direction of the major axes of the ellipses is along the characteristic dimension, and in Fig. 9-11 it is along the vertical dimensions of the bricks.

Once the characteristic dimensions and the tilt have been obtained, distances of the various texture elements can be determined simply by the magnitudes of the characteristic dimensions, since they are inversely proportional to the distance. The slant component of the surface orientation can be computed from these distances. An alternate expression for slant in terms of the gradient of the characteristic dimension δ is given by

$$\tan \sigma = \frac{\nabla \delta}{\delta} \qquad (9\text{-}7)$$

Bajcsy and Liebermann, used the period of a regular texture as the dimension σ, computed from a Fourier analysis of the texture, to compute surface orientation [45, 46]. (Their analysis is, however, not in terms of the use of a characteristic dimension.)

The above method is ineffective if the projection is orthographic—that is, the surface is far from the viewer. In this case a measure based on the foreshortening gradient may be useful. A measure with this property is the *height/width* ratio or the *aspect ratio* of a texture element—for example, the ratio of the major to the minor axes dimensions for the ellipses of Fig. 9-12. The "width" and the "height" of an element are measured along and normal to the characteristic dimension. If the aspect ratio is ϵ, slant is given, as before, by

$$\tan \sigma = \frac{\nabla \epsilon}{\epsilon} \tag{9-8}$$

The above measure assumes uniformity of the texture elements but not their spacing. It may not apply to natural textures, as the sizes of the elements (such as pebbles and grass blades) may change and, further, some elements may be occluded by others. A related technique, using the distribution of the directions of the tangents to the texture element boundaries as a function of the surface slant, is described in [47] and has been applied to simple natural textures.

Kender has explored the constraints imposed on the local surface shape, if certain assumptions are made about the texture property [48]. These constraints are expressed in terms of a *normalized texture property* that must be present in the 3-D scene for an assumed orientation of the surface, and they are expressed as contours in the gradient space, analogous to the reflectance maps of the previous section. The texture properties may be length, slope, density, and so on. Use of the gradient space implies that the projection is orthographic. Also, the texture elements are assumed to be "painted" on the surface, rather than "pointed" away from the surface (for example, trees in a forest texture).

For example, consider a vertical line texture primitive of a given length in an image. The "deprojected" length of this line in the 3-D scene depends on the local orientation of the surface; that is, the deprojected length is fixed for each point (p, q) in the gradient space. The constraints can be expressed as equal-length contours; in this case they consist of vertical parallel lines in the (p, q) plane. Note that this information alone is not sufficient to specify the surface orientation.

Now, if another texture primitive consists of a line of another orientation, a corresponding map can be constructed for this line. If it is further assumed that the two lines have the same length in the scene,

then the surface must lie at the intersection of the contours in the two maps. A third texture property such as a known angle between the two lines (in the scene) is needed to constrain the surface to be one of the two orientations, one being the reflection of the other in the gradient space. Such a relation may come, for example, by assuming that orthogonal lines in the scene are orthogonal on the 3-D surface.

Note that the above analysis only specifies constraints under certain assumptions and is consistent with the analysis of shape from shading of the previous section if intensity is considered as a primitive texture property. The texture properties to use must still be found by the use of *heuristic* rules.

9.6 CONTOUR ANALYSIS

The boundaries of objects themselves convey some 3-D shape information; for example, see Fig. 9-14, where the two figures have similar line configurations but are perceived to be of different shape (a cube and a truncated pyramid). We are also able to infer 3-D shape from 2-D contour lines that are projections of 3-D curves; such representations are common for graphic display of 3-D functions. Techniques for inferring shape from contours basically assume that the observed regularities or near regularities, such as parallelism and symmetry, are not accidental and correspond to similar regularities on the 3-D surface.

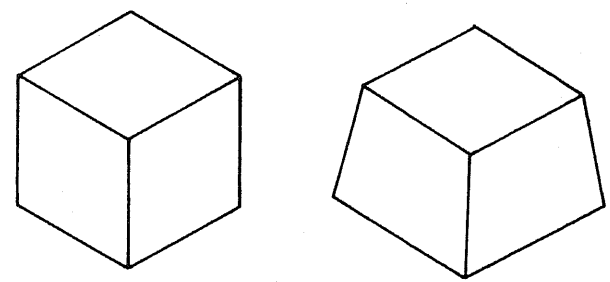

Figure 9-14: Two polyhedral objects of different shape (after Kanade [49])

Kanade has described a method suitable for planar surfaces [49]. His process assumes that parallel lines in the image are also parallel in the scene and that a *skewed symmetry* is the projection of a real

symmetry. A symmetry in 2-D has an axis for which the opposite sides are reflective. In skewed symmetry, the corresponding points need not be perpendicular to the axis but may be at some fixed angle to it. Figure 9-15 shows a figure with skew symmetry and the axes of symmetry. Constraints on the orientation of the plane of this surface can be determined if it is assumed that the axes of the skew symmetry are projections of the orthogonal axes of real symmetry. (The constraints are in fact a hyperbola in the gradient space. A particular pair of orientations may be chosen by picking the points closest to the origin. Details are given in [49]. Stevens has described experiments on human observer's ability to estimate surface orientations from skew symmetry using images as in Fig. 9-15 [44].) The two assumptions used by Kanade's method are also sufficient to explain the perceived shapes of the two objects in Fig. 9-14. A promising start has been made by Lowe and Binford in interpretation of three-dimensional structure from image curves for much more general scenes [50, 51]. These techniques also make use of shadow information to infer heights of objects above ground.

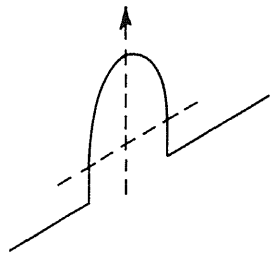

Figure 9-15: A skew-symmetric figure and axes of symmetry

Barrow and Tenenbaum have suggested integrating various analyses, yielding *intrinsic scene characteristics* [40]. These characteristics, for each point in the image, are the distance from the viewer, surface orientation, reflectance, and illumination. Partial object contours can provide local characteristics, which are then propagated inward, and consistency of the intrinsic values is established. This processing should, in turn, lead to the completion of the initial partial contours. The described processing is, again, directly applicable to simple scenes only.

9.7 SUMMARY

In this chapter we have studied techniques for measuring or estimating the three-dimensional positions of the visible surfaces of objects in a scene. Such information is helpful in scene segmentation. Depth measurement by use of stereo views is passive, but the problems of global correspondence need to be solved. Active ranging requires use of special equipment and may not be suitable under all conditions.

Range can also be estimated from monocular images, using variations in intensity, texture gradients, and the contours in the image. Such estimates require assumptions about the scene being viewed. Current techniques are applicable only under special conditions and for simple scenes only.

REFERENCES

[1] M. J. Hannah, *Computer Matching of Areas in Stereo Images*, Stanford Artificial Intelligence Laboratory Memo AIM-239, July 1974 (Ph.D. thesis).

[2] D. B. Gennery, "Modelling the Environment of an Exploring Vehicle by Means of Stereo Vision," Stanford Artificial Intelligence Laboratory Memo AIM-339, June 1980.

[3] D. J. Burr and R. T. Chien, "A System for Stereo Computer Vision with Geometric Models," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Boston, 1977, p. 583.

[4] R. D. Arnold, "Local Context in Matching Edges for Stereo Vision," *Proceedings of Image Understanding Workshop*, Cambridge, Mass., May 1978, pp. 65-72.

[5] H. H. Baker and T. O. Binford, "Depth from Edge and Intensity Based Stereo," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, Canada, August 1981, pp. 631-636.

[6] D. Marr and T. Poggio, "A Computational Theory of Human Stereo Vision," *Proceedings of Royal Society of London*, B204, 1979, pp. 301-328.

[7] W. E. L. Grimson, "Aspects of a Computational Theory of Human Stereo Vision," *Proceedings of Image Understanding Workshop*, University of Maryland, April 1980, pp. 128-149.

[8] S. Ganapathy, "Reconstruction of Scenes Containing Polyhedra from Stereo Pair of Views," Stanford Artificial Intelligence Laboratory Memo AIM-272, December 1975.

[9] S. A. Underwood and C. L. Coates, Jr., "Visual Learning from

Multiple Views," *IEEE Transactions on Computers*, Vol. 24, No. 6, June 1975, pp. 651-661.

[10] R. Shapira and H. Freeman, "Reconstruction of Curved Surface Bodies from a Set of Imperfect *Projections*," *Proceedings of Fifth International Joint Conference on Artificial Intelligence*, Cambridge, Mass., August 1977, pp. 628-634.

[11] D. I. Barnea and H. F. Silverman, "A Class of Algorithms for Fast Digital Image Registration," *IEEE Transactions on Computers*, C-21, 1972, 179-186.

[12] R. Nevatia, "Depth Measurement by Motion Stereo," *Computer Graphics and Image Processing*, Vol. 5, 1976, pp. 203-214.

[13] B. G. Baumgart, "Geometric Modelling for Computer Vision," Stanford Artificial Intelligence Laboratory Memo AIM-249, October 1974.

[14] H. P. Moravec, "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Stanford Artificial Intelligence Laboratory Memo AIM-340, September 1980.

[15] B. Julesz, *Foundations of Cyclopean Perception*, University of Chicago Press, Chicago, 1971.

[16] P. Dev, "Perception of Depth Surfaces in Random-Dot Stereograms A Neutral Model," *International Journal of Man-Machine Studies*, Vol. 7, 1975, pp. 511-528.

[17] R. L. Lillestrand, "Techniques for Change Detection," *IEEE Transactions on Computers*, Vol. 21, July 1972, pp. 654-659.

[18] H. H. Nagel, "Analysis Techniques for Image Sequences," *Proceedings of the Fourth International Joint Conference on Pattern Recognition*, Kyoto, Japan, November 1979, pp. 186-211.

[19] C. L. Fennema and W.B. Thompson, "Velocity Determination in Scenes Containing Several Moving Objects," *Computer Graphics and Image Processing*, Vol. 9, April 1979, pp. 301-315.

[20] W. B. Thompson, "Combining Motion and Contrast for Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 2, No. 6, November 1980, pp. 543-549.

[21] R. Jain, W. N. Martin and J. K. Aggarwal, "Segmentation Through the Detection of Changes Due to Motion," *Computer Graphics and Image Processing*, Vol. 11, September 1979, pp. 13-34.

[22] S. Ullman, *The Interpretation of Visual Motion*, MIT Press, Cambridge, Mass., 1979.

[23] R. F. Rashid, "Towards a System for the Interpretation of Moving Light Displays," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 2, No. 6, November 1980, pp. 524-581.

[24] B. K. P. Horn and B. G. Schunk, "Determining Optical Flow," MIT Artificial Intelligence Laboratory Memo No. 572, Cambridge, Mass., 1980.

[25] K. Prazdny, "Determining the Instantaneous Direction of Motion from Optical Flow Generated by a Curvilinearly Moving Observer," *Proceedings of Image Understanding Workshop*, Washington, D.C., April 1981, pp. 14-21.

[26] G. J. Agin and T. O. Binford, "Computer Description of Curved Objects," *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Stanford, Calif., 1973, pp. 629-640.

[27] Y. Shirai, "Recognition of Polyhedra with a Range Finder," *Pattern Recognition*, Vol. 4, 1972, pp. 243-250.

[28] R. Nevatia, *Computer Analysis of Scenes of 3-Dimensional Curved Objects*, Birkhauser-Verlag, Basel, Switzerland, 1976.

[29] P. M. Will and K. S. Pennington, "Grid Coding: A Preprocessing Technique for Robot and Machine Vision," *Artificial Intelligence*, Vol. 2, 1971, pp. 319-329.

[30] R. A. Lewis and A. R. Johnson, "A Scanning Laser Rangefinder for a Robotic Vehicle," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Cambridge, Mass., August 1977, pp. 762-768.

[31] D. Nitzan, A. E. Brain, and R. O. Duda, "The Measurement and Use of Registered Reflectance and Range Data in Analysis," *Proceedings of IEEE*, Vol. 65, February 1977, pp. 206-220.

[32] S. Inokuchi and R. Nevatia, "Boundary Detection in Range Pictures," *Proceedings of Fifth International Conference on Pattern Recognition*, Miami, November 1980, pp. 1301-1303.

[33] R.O. Duda, D. Nitzan, and P. Barrett, "Use of Range and Reflectance Data to Find Planar Surface Regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 1, No. 3, July 1979, pp. 259-271.

[34] K. Sugihara, "Range-Data Analysis Guided by a Junction Dictionary," *Artificial Intelligence*, Vol. 12, 1979, pp. 41-69.

[35] M. Oshima and Y. Shirai, "A Scene Description Method Using Three-dimensional Information," *Pattern Recognition*, 1979, pp. 9-17.

[36] B. Horn, "Obtaining Shape from Shading Information," in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, New York, 1975.

[37] T. Rindfleisch, "Photometric Method for Lunar Topography," *Photogrammetric Engineering*, Vol. 32, 1966, pp. 262-276.

[38] B. Horn, "Understanding Image Intensities," *Artificial Intelligence*, Vol. 8, 1977, pp. 201-231.

[39] R. J. Woodham, "A Cooperative Algorithm for Determining Surface Orientation from a Single View," *Proceedings of the Fifth*

International Joint Conference on Artificial Intelligence, Cambridge, Mass., 1977, pp. 635-641.

[40] H. Barrow and J.M. Tenenbaum, "Recovery of Intrinsic Scene Characteristics from Images," in *Computer Vision Systems*, A. Hanson and E. Riseman, (eds.), Academic Press, New York, 1978, pp. 3-26.

[41] R. J. Woodham, "Reflectance Map Techniques for Analyzing Surface Defects in Metal Castings" MIT Artificial Intelligence Laboratory Memo AI-TR-457, June 1978.

[42] B. K.P. Horn and B. L. Bachman, "Using Synthetic Images to Register Real Images with Surface Models," *Communications of ACM*, Vol. 21, No. 11, November 1978, pp. 914-924.

[43] J. J. Gibson, *The Perception of the Visual World*, Houghton Mifflin, Boston, Mass., 1950.

[44] K. Stevens, "Surface Perception from Local Analysis of Texture and Contour," MIT Artificial Intelligence Laboratory Memo AI-TR-512, February 1980.

[45] R. Bajcsy, "Computer Identification of Visual Surface," *Computer Graphics and Image Processing*, Vol. 2, No. 2, 1973, pp. 118-130.

[46] R. Bajcsy and L. Liebermann, "Texture Gradient as a Depth Cue," *Computer Graphics and Image Processing*, Vol. 5, No. 1, March 1976, pp. 52-67.

[47] A. P. Witkin, "A Statistical Technique for Recovering Surface Orientation from Texture in Natural Imagery," *Proceedings of the First Annual National Conference on Artificial Intelligence*, Stanford, Calif., August 1980, pp. 1-3.

[48] J. R. Kender, "Shape from Texture," Carnegie-Mellon University, Computer Science Technical Report CMU-CS-81-102, Pittsburgh, November 1980.

[49] T. Kanade, "Recovery of the Three-Dimensional Shape of an Object from a Single View," Carnegie-Mellon University, Computer Science Report CMU-CS-79-153, Pittsburgh, October 1979.

[50] D. G. Lowe and T. O. Binford, "The Interprettion of Three-Dimensional Structure from Images," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, Canada, August 1981, pp. 613-618.

[51] T. O. Binford, "Inferring Surfaces from Images," *Artificial Intelligence*, Vol. 17, 1981, pp. 205-244.

10

SYSTEMS AND APPLICATIONS

Visual perception systems may be "general purpose" or tailored for special tasks. A general-purpose system is expected to have capabilities similar to the human visual system and handle a wide variety of scenes under a variety of viewing conditions. Under normal viewing conditions, humans may expect of certain objects to be present in the scene, but our ability to perceive seems to be almost as good when we are presented with a randomly chosen photograph. We are able to generate high-quality descriptions of unfamiliar objects, such as pictures of new planets, or photomicrographs of molecules. General-purpose vision may be defined to have capabilities similar to human perception; a more fundamental definition is difficult, owing to the inherent ambiguity of the images.

Our understanding of the perceptual processes needed to achieve general vision is poor, and the performance of the techniques discussed in the previous chapters is low in comparison to human performance. Fortunately, a great many applications of practical importance do not require this generality, as the domain of objects is often small, and significant knowledge of the scene is available a priori. Special-purpose *knowledge-based systems* aim to maximize the utilization of such knowledge.

In this chapter we examine some requirements for a general-purpose system and describe some knowledge-based systems with applications.