Second Edition

ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE

Volume 2
M-Z

Awarded

American Library Association's

Outstanding Reference Source

Association of American Publishers Award

Best New Professional and Scholarly Publication

Stuart C. Shapiro
Editor-in-Chief

V

VISUAL MOTION ANALYSIS

When an observer moves relative to the environment, the two-dimensional (2-D) image that is projected onto the eye undergoes complex changes. These changes however, contain information regarding the relative 3-D motion and the structure of the scene in view.

There exist several representations for the pattern of movement of features in the image, containing different amounts of information related to 3-D motion and shape. The ones most studied are *optical flow, normal optical flow,* and *discrete displacements.*

OPTICAL FLOW

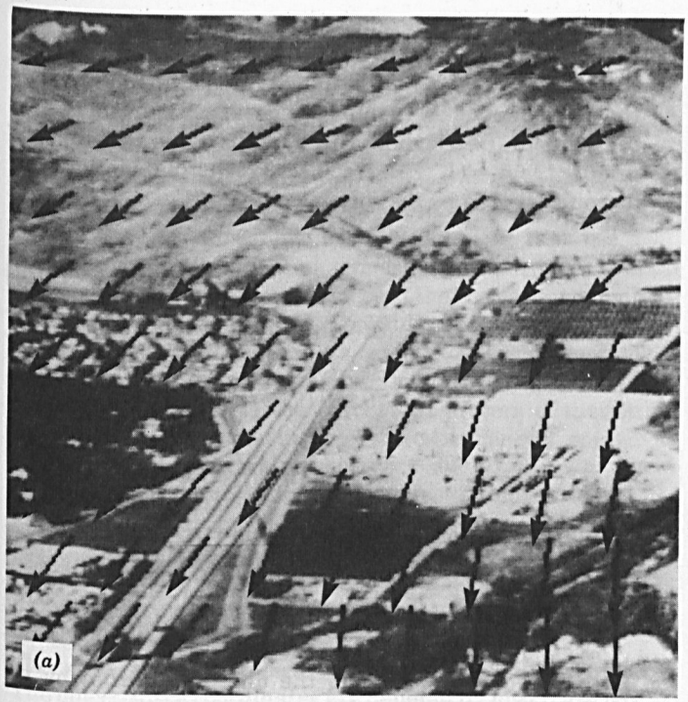
Optical flow (Gibson, 1950) can be represented by a 2-D field of velocity vectors as shown in Figure 1. In Figure 1a the optical flow is generated by the movement of an observer relative to a stationary environment. The "observer" is a camera mounted on an airplane that is flying over terrain. A single snapshot from a sequence of images is shown with reduced contrast. The black vectors superimposed on the image represent the optical flow, or velocity field. The direction and length of these vectors indicate the direction and speed of movement of features across the image as the airplane flies along. Optical flow is also generated by the motion of objects in the environment. Figure 1b shows three views of a three-dimensional (3-D) wire-frame object that is rotating about a central vertical axis. Figure 1c shows a snapshot of the object at a particular moment in time, with vectors superimposed that indicate the velocities of individual points on the object.

The analysis of the optical flow can be divided into two parts: The first is the measurement of optical flow from the changing image, and the second is the use of optical flow to recover important properties of the environment. The motion of features in the image is not provided to the visual system directly but must be inferred from the changing pattern of intensity that reaches the eye. Variations in the measured optical flow across the image (also known as motion parallax) can then be used to recover the movement of the observer, the 3-D shape of visible surfaces, and the locations of object boundaries. For example, from a sequence of optical flows such as that shown in Figure 1a, it is possible to recover the motion of the airplane relative to the ground. The variation in speed of movement of points on the wire-frame object of Figure 1c allows the recovery of its 3-D structure from the changing 2-D projection. Sharp changes in the optical flow field indicate the presence of object boundaries in the scene.

Computational studies offer a broad range of methods for measuring optical flow (for reviews, see Thompson and Barnard, 1981; Ullman, 1981; Ballard and Brown, 1982; Hildreth, 1984). Some methods compute the instantaneous optical-flow field directly. Methods for measuring motion also differ in the stage of image processing at which movement is first analyzed. For example, some infer movement directly from changes in the image intensities, and others first filter the image, or extract features such as edges. The range of techniques for motion measurement are reflected in a broad range of application domains, from the simple tracking of objects along a conveyor belt in an industrial setting to the analysis of more complex motions such as that of clouds in satellite weather data, heart walls in x-ray images, or cells in cell cultures. The analysis of optical flow is also becoming essential in autonomous navigation (see ROBOTS, MOBILE) and robotic assembly (see MANUFACTURING, AI IN; ROBOTICS).

The measurement of optical flow poses two fundamental problems for computer-vision systems. First, the changing pattern of image intensity provides only partial information about the true motion of features in the image due to a problem often referred to as the aperture problem. Second, when the general motion of objects is allowed, there does not exist a unique optical-flow field that is consistent with the changing image. In theory, there exist infinite possible interpretations of the motion of features in the image. Additional constraint is required to identify the most plausible interpretation from a physical viewpoint.

The aperture problem is illustrated in Figure 2. Suppose that the movement of features in the image were first detected using operations that examine only a limited area of the image. Such operations can provide only partial information about the true motion of features in the image (Thompson and Barnard, 1981; Ullman, 1981; Ballard and Brown, 1982; Hildreth, 1984; Lawton, 1983; Horn and Schunck, 1981). In Figure 2a the extended edge **E** moves across the image, and its movement is observed through a window defined by the circular aperture **A**. Through this window, it is only possible to observe the movement of the edge in the direction perpendicular to its orientation. The component of motion along the orientation of the edge is invisible through this limited aperture. Thus, it is not possible to distinguish between motions in the directions b, c, and d. This property is true of any motion detection operation that examines only a limited area of the image. As a consequence of the aperture problem, the measurement of optical flow requires two stages of analysis: The first measures components of motion in the direction perpendicular to the orientation of image features; the second combines these components of motion to compute the full 2-D pattern of movement in the image. In Figure 2b a circle undergoes pure translation to the right. The arrows along the contour represent the perpendicular components of velocity that can be measured directly from the changing image. These component measurements each provide some constraint on the possible motion of the circle. Its true motion, however, can be determined only by combining the constraints imposed by these component measurements. The movement of some features such as corners or small patches and spots can be



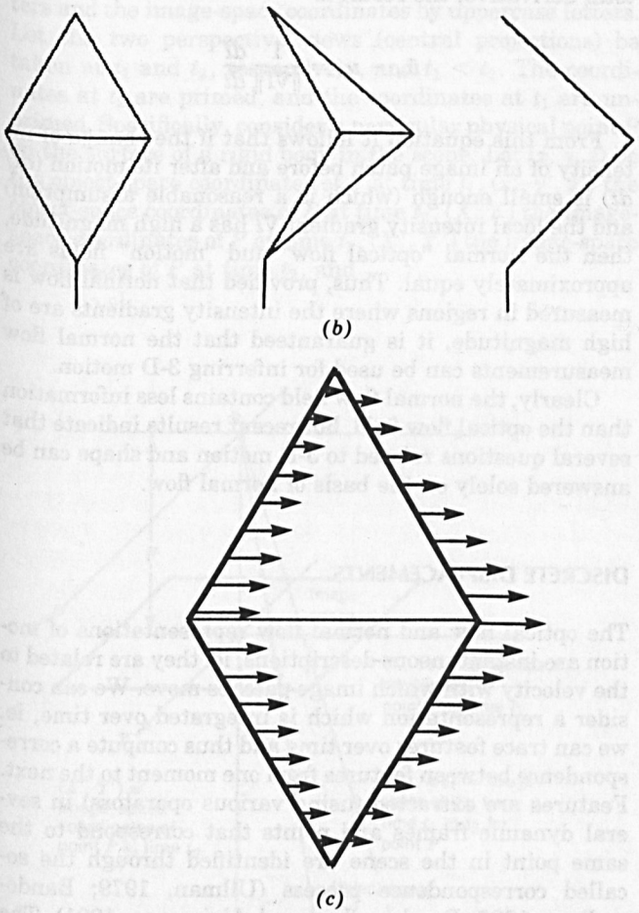


Figure 1. (*a*) Optical-flow field, represented by black arrows, is superimposed on a natural image that was taken from an airplane flying over terrain. (*b*) Three views of a wire-frame object rotating about a central vertical axis. (*c*) Projected pattern of velocities of individual points on the object are shown superimposed on a snapshot of the object in motion (an orthographic, or parallel projection is used).

measured unambiguously in the changing image. Several methods for measuring motion rely on the tracking of such isolated features (Thompson and Barnard, 1981; Ullman, 1981; Ballard and Brown, 1982; Lawton, 1983). In general, however, the first measurements of movement provide only partial information about the true movement of features in the image and must be combined to compute the full optical-flow field.

The measurement of movement is difficult because in theory, there are infinitely many patterns of motion that are consistent with a given changing image. For example, in Figure 2c, the contour **C** rotates, translates, and deforms to yield the contour **C′** at some other time. The true motion of the point *p* is ambiguous. Additional constraint is required to identify a single pattern of motion. Many physical assumptions could provide this additional constraint. One possibility is the assumption of pure translation. That is, it is assumed that velocity is constant over small areas of the image. This assumption has been used both in computer-vision studies and in biological models of motion measurement (Thompson and Barnard, 1981;

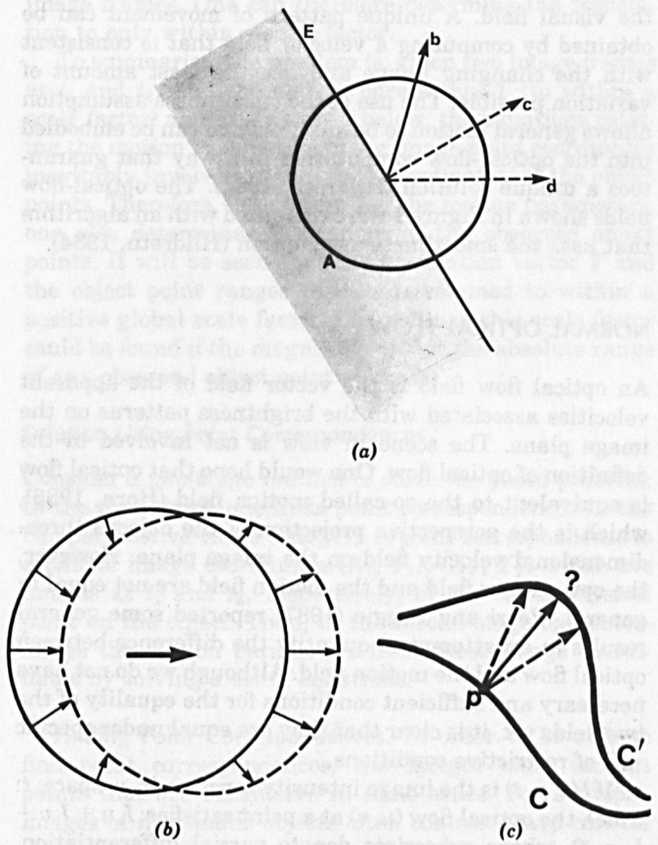


Figure 2. (*a*) Operation that examines the moving edge **E** through the limited aperture **A** can compute only the component of motion *c* in the direction perpendicular to the orientation of the edge. The true motion of the edge is ambiguous. (*b*) A circle undergoes pure translation to the right. The arrows along the circle represent the perpendicular components of motion that can be measured directly from the changing image. (*c*) A contour **C** rotates, translates, and deforms to yield the contour **C′**. The motion of the point *p* is ambiguous.

Ullman, 1981; Ballard and Brown, 1982; Hildreth, 1984; Lawton, 1983; Nakayama, 1985). Methods that assume pure translation are useful for detecting sudden movements and tracking objects across the visual field. These methods have led to fast algorithms for computing a rough estimate of the motion of objects, which is often sufficient in applications of motion analysis. Tasks such as the recovery of 3-D structure from motion require a more detailed measurement of relative motion in the image. The analysis of variations in motion such as those illustrated in Figure 2c requires the use of a more general physical assumption.

Other computational studies have assumed that velocity varies smoothly across the image (Hildreth, 1984; Horn and Schunck, 1981). This is motivated by the assumption that physical surfaces are generally smooth. Variations in the structure of a surface are usually small compared with the distance of the surface from the viewer. When surfaces move, nearby points tend to move with similar velocities. There exist discontinuities in movement at object boundaries, but most of the image is the projection of relatively smooth surfaces. Thus, it is assumed that image velocities vary smoothly over most of the visual field. A unique pattern of movement can be obtained by computing a velocity field that is consistent with the changing image and has the least amount of variation possible. The use of the smoothness assumption allows general motion to be analyzed and can be embodied into the optical-flow computation in a way that guarantees a unique solution (Hildreth, 1984). The optical-flow fields shown in Figure 1 were computed with an algorithm that uses the smoothness assumption (Hildreth, 1984).

NORMAL OPTICAL FLOW

An optical flow field is the vector field of the apparent velocities associated with the brightness patterns on the image plane. The scene in view is not involved in the definition of optical flow. One would hope that optical flow is equivalent to the so-called motion field (Horn, 1986), which is the perspective projection of the object's three-dimensional velocity field on the image plane. However, the optical flow field and the motion field are not equal in general. Verri and Poggio (1987) reported some general results in an attempt to quantify the difference between optical flow and the motion field. Although we do not have necessary and sufficient conditions for the equality of the two fields yet, it is clear that they are equal under specific sets of restrictive conditions.

If $I(x, y, t)$ is the image intensity function (x, y: space; t: time), the optical flow (u, v) at a point satisfies: $I_x u + I_y v + I_t = 0$, where subscripts denote partial differentiation. This equation can be written as $(I_x, I_y)\,(u, v) = I_t$, indicating that the projection of the optical flow (u, v) along the direction (I_x, I_y) is known. This is what is called the *normal optical flow*.

Clearly, estimating normal flow is much easier than estimating the actual optical flow. But then, how is normal flow related to the three-dimensional motion field? Is the normal optical flow field equal to the normal motion field, and under what conditions?

Let $I(x, y, t)$ denote the image intensity, and consider the optical flow field $(u, v) = \vec{v}$ and the motion field $\vec{\bar{v}} = (\bar{u}, \bar{v})$ at a point (x, y) where the local (normalized) intensity gradient is $\vec{n} = (I_x, I_y)/\sqrt{I_x^2 + I_y^2}$. The normal motion field at point (x, y) is by definition

$$\bar{u}_n = \vec{\bar{v}} \cdot \vec{n} \quad \text{or}$$

$$\bar{u}_n = \left(\frac{dx}{dt}, \frac{dy}{dt}\right) \cdot \frac{(I_x, I_y)}{\sqrt{I_x^2 + I_y^2}} \quad \text{or}$$

$$\bar{u}_n = \left(\frac{dx}{dt}, \frac{dy}{dt}\right) \cdot \frac{\nabla I}{\|\nabla I\|} \quad \text{or}$$

$$\bar{u}_n = \frac{1}{\|\nabla I\|}\left(I_x \frac{dx}{dt} + I_y \frac{dy}{dt}\right)$$

Similarly, the normal optical flow is

$$u_n = -\frac{1}{\|\nabla I\|} I_t$$

Thus, when approximating the differential dI/dt by its total derivative, the result is

$$\bar{u}_n - u_n = \frac{1}{\|\nabla I\|} \frac{dI}{dt}$$

From this equation it follows that if the change of intensity of an image patch before and after its motion (dI/dt) is small enough (which is a reasonable assumption) and the local intensity gradient ∇I has a high magnitude, then the normal "optical flow" and "motion" fields are approximately equal. Thus, provided that normal flow is measured in regions where the intensity gradients are of high magnitude, it is guaranteed that the normal flow measurements can be used for inferring 3-D motion.

Clearly, the normal flow field contains less information than the optical flow field, but recent results indicate that several questions related to 3-D motion and shape can be answered solely on the basis of normal flow.

DISCRETE DISPLACEMENTS

The optical flow and normal flow representations of motion are instantaneous descriptions, ie, they are related to the velocity with which image patches move. We can consider a representation which is integrated over time, ie, we can trace features over time and thus compute a correspondence between features from one moment to the next. Features are extracted (using various operators) in several dynamic frames and points that correspond to the same point in the scene are identified through the so-called correspondence process (Ullman, 1979; Bandopadhay, 1986; Bandopadhay and Aloimonos, 1991). The latter sections will describe various approaches to the determination of three-dimensional motion of a rigid body based on time-sequential perspective views.

Determining the relative motion between an observer and his environment is a major problem in computer vision. Its applications include mobile-robot (see ROBOTS, MOBILE) navigation and monitoring dynamic industrial pro-

cesses. For background material, the reader is referred to the two edited volumes of Huang (1981, 1983), the pioneering and influential book of Ullman (1979), several special journal issues and proceedings of several workshops on motion (see General References).

The next three sections describe methods that use a monocular two-dimensional sensor (such as a television camera); then methods are discussed that use a stereo pair of sensors. Finally, there is a brief discussion on numerical accuracy, multiple objects, nonrigid objects, motion prediction, and high level motion understanding. We consider as the inputs to the perceptual process of motion analysis discrete displacements (correspondences), optical flow, and normal optical flow.

TWO-VIEW MOTION ANALYSIS USING FEATURE CORRESPONDENCE

Problem Statement

The basic geometry of the problem is sketched in Figure 3. The object-space coordinates are denoted by lowercase letters and the image-space coordinates by uppercase letters. Let the two perspective views (central projections) be taken at t_1 and t_2, respectively, and $t_1 < t_2$. The coordinates at t_2 are primed, and the coordinates at t_1 are unprimed. Specifically, consider a particular physical point P on the surface of a rigid body in the scene. Let (x, y, z) be the object-space coordinates of P at time t_1, (x', y', z') the object-space coordinates of P at time t_2, (X, Y) the image-space coordinates of P at time t_1, (X', Y') the image-space coordinates of P at time t_2, and

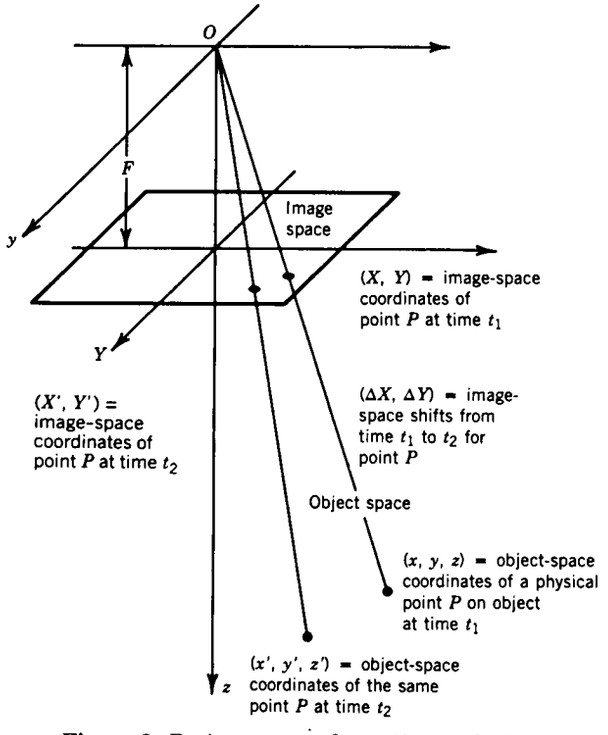


Figure 3. Basic geometry for motion analysis.

$$\Delta X \triangleq X' - X \qquad \Delta Y \triangleq Y' - Y \tag{1}$$

the image-space shifts (or displacements) of P from t_1 to t_2.

It is well known from kinematics that the object coordinates of P at time instants t_1 and t_2 are related by

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R \begin{bmatrix} x \\ y \\ z \end{bmatrix} + T = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} \tag{2}$$

where R represents a rotation and T a translation. To make the representation unique, the rotation is specified around an axis passing through the origin of the coordinate system. Let $\hat{n} = (n_1, n_2, n_3)$ be a unit vector along the axis of rotation and θ be the angle of rotation from t_1 to t_2. Then the elements of R can be expressed in terms of n_1, n_2, n_3, and θ. Since $n_1^2 + n_2^2 + n_3^2 = 1$, there are six motion parameters to be determined: n_1, n_2, θ, Δx, Δy, and Δz. However, from the two perspective views, it is impossible to determine the magnitude of the translation, ie, if the object size and position as well as the translation are scaled by the same factor, one gets exactly the same two image frames. One can therefore determine the translation to only within a scale factor.

To summarize, the problem is: given two image frames at t_1 and t_2, find the motion parameters T (to within a scale factor) and R. As shown below, the equations relating the motion parameters to the image-point coordinates inevitably involve the ranges (z coordinates) of the object points. Therefore, in determining the motion parameters, one also determines the ranges of the observed object points. It will be seen that the translation vector T and the object point ranges can be determined to within a positive global scale factor. The value of this scale factor could be found if the magnitude of T or the absolute range of any observed object point is known.

Solution Using Point Correspondences

Consider a two-stage method to solve the posed problem. In the first stage, one finds point correspondences in the two perspective views (images). A point correspondence is a pair of image coordinates (X_i, Y_i), (X_i', Y_i') which are images at t_1 and t_2, respectively, of the same physical point on the object. Then, in the second stage one determines the motion parameters from these image coordinates by solving a set of equations.

Finding Point Correspondences. In order to be able to find point correspondences, the images must contain points that are distinctive in some sense. For example, images of man-made objects often contain sharp corners that are relatively easy to extract (Fang and Huang, 1982). More generally, image points where the local gray-level variations (defined in some way) are maximum can be used (Moravec, 1980). Other important approaches include Nagel (1983) and Kories and Zimmermann (1986).

In any case, in each of the two images a large number of distinctive points are extracted. Then one tries to match the two point patterns in the two images using spatial structures of the patterns (Fang and Huang, 1984). The matching will be successful only if the amount of rotation

(θ) is relatively small (so that the perspective distortion is small). For example, in Fang and Huang (1982), good matching results are obtained if $\theta < 5°$. This restriction may be relaxed if there is some a priori information about the object (Gu and co-workers, 1984).

Basic Equations. From Figure 3 there is the following relationship between the image-space and the object-space coordinates:

$$X = F\frac{x}{z} \qquad Y = \frac{y}{z} \tag{3}$$

For simplicity, assume throughout that $F = 1$. The motion is described by Eq. 2. From Eqs. 2 and 3,

$$X' = \frac{(r_{11}X + r_{12}Y + r_{13})z + \Delta x}{(r_{31}X + r_{32}Y + r_{33})z + \Delta z}$$

$$Y' = \frac{(r_{21}X + r_{22}Y + r_{23})z + \Delta z}{(r_{31}X + r_{32}Y + r_{33})z + \Delta z} \tag{4}$$

where the r_{ij} can be expressed in terms of n_1, n_2, n_3, and θ. By elimination of z from Eq. 4,

$$(\Delta x - X'\,\Delta z)\{y'(r_{31}X + r_{32}Y + r_{33}) - (r_{21}X + r_{22}Y + r_{23})\}$$
$$= (\Delta y - Y'\,\Delta z)\{X'(r_{31}X + r_{32}Y + r_{33}) - (r_{11}X + r_{12}Y + r_{13})\} \tag{5}$$

Also,

$$z = \frac{\Delta x - X'\,\Delta z}{X'(r_{31}X + r_{32}Y + r_{33}) - (r_{11}X + r_{12}Y + r_{13})}$$
$$= \frac{\Delta y - X'\,\Delta z}{Y'(r_{31}X + r_{32}Y + r_{33}) - (r_{21}X + r_{22}Y + r_{23})} \tag{6}$$

Equation 5 is nonlinear in the six unknowns: Δx, Δy, Δz, n_1, n_2, and θ. Also, it is homogeneous in Δx, Δy, and Δz. Therefore, as mentioned earlier, one can only hope to find T to within a scale factor. After finding T (to within a scale factor) and R, one can find z_i for each observed point to within the same scale factor using Eq. 6.

To fix ideas, let the translation sought after be the unit translation vector

$$\hat{T} = (\Delta\hat{x}, \Delta\hat{y}, \Delta\hat{z}) \triangleq \frac{1}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}\,T \tag{7}$$

Then, Eq. 5 can be considered as a nonlinear equation in the five unknowns: $\Delta\hat{x}$, $\Delta\hat{y}$, n_1, n_2, and θ. Thus, with 5-point correspondence, there are five equations with five unknowns. Well-known iterative techniques can then be used to find solutions. In practice, because of noise in the image data, one tries to find more than 5-point correspondences and seek a least-squares solution.

Alternative Formulation. The motion-parameter Eq. 5 was derived by eliminating z in Eq. 4. Alternatively, one can formulate equations in terms of the z coordinates of the points under consideration without containing any motion parameters (Mitchie and Aggarwal, 1985). This can be done by using the principle of distance conservation for a rigid body. Assume N point correspondences are given:

$$(X_i, Y_i) \leftrightarrow (X_i', Y_i') \; i = 1, 2, \ldots, N$$

And let (x_i, y_i, z_i) and (x_i', y_i', z_i') be the 3-D coordinates of the ith point at t_1 and t_2, respectively. Then, one has

$$(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 = (x_i' - x_j')^2 + (y_i' - y_j')^2 + (z_i' - z_j')^2 \tag{8}$$

and from Eq. 3

$$(z_iX_i - z_jX_j)^2 + (z_iY_i - z_jY_j)^2 + (z_i - z_j)^2 = (z_i'X_i' - z_j'X_j')^2 + (z_i'Y_i' - z_j'Y_j')^2 + (z_i' - z_j')^2 \tag{9}$$

For each pair of points, one Eq. 9 can be written. Thus, with five-point correspondences, one can write ten equations that (if $z_1 = 1$) contain nine unknowns: $z_2, \ldots, z_5$, $z_1', \ldots, z_5'$. A least-squares solution for these unknowns can be found using iterative methods. Then the motion parameters are found by solving Eq. 2. Several methods for carrying out the last step are discussed under Motion from 3-D Feature Correspondences.

Disadvantage of Solving Nonlinear Equations. To find a least-squares solution of a small set of nonlinear equations 5 or 9 using iterative methods is not computationally expensive. However, unless there is a good initial-guess solution, the iteration may not coverge or it may converge to a local but not global minimum. Furthermore, with nonlinear equations it is very difficult to analyze the question of solution uniqueness.

In fact, it is an open theoretical question: what is the minimum number of point correspondences that will ensure a unique solution for the five motion parameters $\Delta\hat{x}$, $\Delta\hat{y}$, n_1, n_2, and θ? With 5-point correspondences the number of equations become equal to or larger than the number of unknowns. However, since the equations are nonlinear, one would expect that the solution may generally not be unique. This has indeed been verified by computer simulations in which global searches were made. The results of such simulations indicated that with 5-point correspondences there may be more than one solution; with 6-or-more-point correspondences the solution is generally unique. It is to be noted that in the case of 5-point correspondences, even though the solution may not be unique, if the iteration is started at a guess solution that is close to the true solution, one will most likely converge to it.

The conclusion is that the approach of solving nonlinear equations is viable if there is a good initial-guess solution. Otherwise, a better alternative is described in the next section: A linear algorithm that requires 8-or-more-point correspondences.

A Linear Algorithm. It turns out that by introduction of appropriate intermediate variables (which are functions of the motion parameters), Eq. 5 becomes linear (Longuet-Higgins, 1981; Tsai and Huang, 1984). Define

$$E = \begin{bmatrix} e_1 & e_2 & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix} = GR \tag{10}$$

where

$$G = \begin{bmatrix} 0 & -\Delta\hat{z} & \Delta\hat{y} \\ \Delta\hat{z} & 0 & -\Delta\hat{x} \\ -\Delta y & \Delta\hat{x} & 0 \end{bmatrix} \quad \text{(skew symmetric)} \tag{11}$$

$\hat{T} = (\Delta\hat{x}, \Delta\hat{y}, \Delta\hat{z})$ is the unit translation vector defined in Eq. 7, and R is the orthonormal rotation matrix. Then Eq. 5 becomes

$$[X' \quad Y' \quad 1]E\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = 0 \tag{12}$$

which is linear and homogeneous in the nine new unknowns: $e_1, \ldots, e_9$. The algorithm consists of two steps:

Step 1. From 8 or more point correspondences determine E to within an unknown scale factor k.

Step 2. decompose kE to obtain R and $\hat{T}$

Step 1 is relatively simple; it amounts to finding the least-squares solution of a set of linear equations 12. Step 2 is more complicated and is not discussed here. Several algorithms are given in other sources (Longuet-Higgins, 1981; Tsai and Huang, 1984; Yen and Huang, 1983; Zhuang and co-workers, 1986; Huang, 1985). It can be shown that, except for degenerate cases, 8 or more point correspondences yield a unique solution for R and $\hat{T}$ (Zhuang and co-workers, 1986; Huang, 1985; Longuet-Higgins, 1984).

Planar Patch Case. In many applications the points observed may all lie on a rigid planar patch in 3-D. In this case the linear algorithm shown above breaks down. One can go back to use the nonlinear equations 5 or 9. However, it turns out that a more computationally efficient, and in fact linear, algorithm exists for the planar patch case (Tsai and Huang, 1981, 1984; Tsai and co-workers, 1983). This linear algorithm, described below, also throws light on the uniqueness question for the planar case.

Let the 3-D points observed all lie on a plane whose equation at t_1 is

$$ax + by + cz = 1$$

or

$$[a, b, c]\begin{bmatrix} x \\ y \\ z \end{bmatrix} = 1 \tag{13}$$

Later, the notation $g = [a, b, c]^t$ (the superscript t denotes transportation) is used. From Eqs. 2 and 13

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R\begin{bmatrix} x \\ y \\ z \end{bmatrix} + T = R\begin{bmatrix} x \\ y \\ z \end{bmatrix} + T[a, b, c]\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$= (R + T[a, b, c])\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

or

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = A\begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{14}$$

where

$$A = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix} = R + \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix}[a, b, c] = R + Tg^t \tag{15}$$

From Eqs. 3 and 14

$$X' = \frac{a_1X + a_2Y + a_3}{a_7X + a_8Y + a_9} \qquad Y' = \frac{a_4x + a_5Y + a_6}{a_7X + a_8Y + a_9} \tag{16}$$

Some other useful formulas are, from Eq. 13,

$$\frac{1}{z} = aX + bY + c \tag{17}$$

and, from Eq. 14,

$$\frac{z'}{z} = a_7x + a_8Y + a_9 \tag{18}$$

The two-step linear algorithm is as follows:

Step 1. From 4 or more point correspondences, a set of linear homogeneous equations 16 are solved to find A to within a scale factor.

Step 2. From A R, wT, and g/w, are determined where w is a positive scale factor.

Step 1 involves basically finding the least-squares solution of a set of linear equations. Step 2 is more complicated; an algorithm using singular-value decomposition is described in Tsai and co-workers (1983). It can be shown that, except for degenerate cases, given 4 or more point correspondences, there are generally two solutions for R, $\hat{T}$, and g. With 4 or more point correspondences over three views, the solution becomes unique (Tsai and Huang, 1985).

Solution Using Straight-Line Correspondences

In the presence of image noise and/or due to the spatial sampling, the coordinates of feature points cannot be determined accurately. This may make the estimation of motion parameters unreliable. Usually, it is easier to detect and determine the location of straight edges than feature points (Yen and Huang, 1986; Liu and Huang, 1986). Therefore, the question arises: can one estimate 3-D motion parameters by using straight-line correspondences?

Finding Straight-Line Correspondences. Images of man-made objects often contain straight edges. These straight edges can be detected using edge point detectors (such as the Sobel operator) followed by Hough transform (qv) (Duda and Hart, 1973). One first detects straight edges in both image frames and then uses structural information

to match the two straight-line patterns. The algorithm of Cheng and Huang (1984) can be used to do the matching if the motion from t_1 and t_2 is small.

Two-View Nonuniqueness. By a straight-line correspondence over two frames, one knows the equations in the image plane at t_1 and t_2 of a 3-D line on the object:

$$t_1: \quad \alpha X + \beta Y = 1 \tag{19}$$

$$t_2: \quad \alpha' X + \beta' Y = 1 \tag{20}$$

where $(\alpha, \beta) \leftrightarrow (\alpha', \beta')$. Note that one does not assume any point correspondences on the two lines. Unfortunately, a little reflection convinces one that no matter how many straight-line correspondences are known over two frames, it is impossible to determine R and $\hat{T}$ uniquely. Heuristically, one can argue as follows: From the imaging system geometry expressions for α' and β' can be derived in terms of R, $\hat{T}$, α, β, and some additional parameters that pin down the position of the 3-D line at t_1. Given the 2-D image of a 3-D line, one needs two additional parameters (γ and δ, say) to determine the 3-D position of the line. Thus,

$$\alpha' = \alpha'(R, \hat{T}\, \alpha, \beta, \gamma, \delta) \tag{21}$$

$$\beta' = \beta'(R, \hat{T}\, \alpha, \beta, \gamma, \delta) \tag{22}$$

Each new straight-line correspondence gives two new equations 21 and 22 but also two new unknowns, γ and δ. Therefore, the number of equations is always smaller than the number of unknowns by five (the five motion parameters).

Three-View Case. With straight-line correspondences over three image frames (at $t_1 < t_2 < t_3$), it is possible to determine the motion parameters R_{12}, $\hat{T}_{12}$, (from t_1 to t_2) and R_{23}, $\hat{T}_{23}$ (from t_2 to t_3). An equation involving R_{12} and R_{23} can be obtained as follows. Let the equations in the image plane at t_1, t_2, and t_3 of a 3-D straight line be given by Eqs. 19, 20, and

$$t_3 \quad \alpha'' X + \beta'' Y = 1 \tag{23}$$

Equation 19 implies with the help of Eq. 3, that at t_1, the 3-D straight line lies in the plane

$$\alpha x + \beta y - z = 0, \tag{24}$$

which has a normal

$$q = (\alpha, \beta, -1) \tag{25}$$

Similarly, at t_2 and t_3, respectively, there are the normals

$$q' = (\alpha', \beta', -1) \tag{26}$$

$$q'' = (\alpha'', \beta'', -1) \tag{27}$$

Then, it can be shown that the tree vectors q', $R_{12}q$, and $R^{-1}_{23}q''$ are coplanar. Thus

$$q' \cdot (R_{12}q \times R^{-1}_{23}q'') = 0 \tag{28}$$

Here a three-element array is considered as either a vector or a column matrix from context. Equation 28 is nonlinear in the six unknown motion parameters (three from each rotation matrix). It has been found empirically that given seven or more straight-line correspondences over three frames, one can determine a unique solution to R_{12} and R_{23} by finding the least-squares solution of the set of nonlinear Eq. 28 using iterative methods. Once the rotations are found, the unit translation vectors can be obtained by solving linear equations. For a complete analysis, see Spetsakis and Aloimonos (1990).

An alternative treatment of the line correspondence case was given by Mitiche, Seida, and Aggarwal (Mitiche and co-workers, 1986).

Solution Using Planar Curve Correspondences

In some cases it may be possible to track the projection of a planar contour (eg, the boundary of a face of a polyhedron) from one image frame to the next. The change in the shape of the 2-D region (in image plane) bounded by the contour contains information on the 3-D motion parameters as well as the orientation of the plane in 3-D. More generally, if more than one region can be tracked the change in the relative positions of these regions (in image plane) can also be utilized. Gambotto and Huang (1984) have shown in a simple example how this region-based method can be used in motion analysis. However, a general methodology, even for the one-region situation, is yet to be developed. In the following, two special cases (one-region) are described.

Small-Motion Case. Kanatani (1985) has suggested a method using line (or surface) integrals. It is assumed that the amount of motion from t_1 to t_2 is small. Then

$$R \approx \begin{bmatrix} 1 & -\phi_3 & \phi_2 \\ \phi_3 & 1 & -\phi_1 \\ -\phi_2 & \phi_1 & 1 \end{bmatrix} \tag{29}$$

where

$$\phi_i = n_i \theta \tag{30}$$

Let C_1 and C_2 be the images at t_1 and t_2, respectively, of a 3-D planar contour. The equation of the plane at t_1 is

$$ax + by + cz = 1 \tag{13}$$

Choose a function $F(X, Y)$ (eg, $F = X^2$), and compute

$$I(t_1) = \int_{C_1} F(X, Y)\, ds \tag{31}$$

$$I(t_2) = \int_{C_2} F(X, Y)\, ds \tag{32}$$

where

$$ds = \sqrt{dX^2 + dY^2} \tag{33}$$

Then, it can be shown that

$$\Delta I \triangleq I(t_2) - I(t_2) \approx K_1 \Delta x + K_2 \Delta y + K_3 \Delta z + K_4\phi_1 + K_5\phi_2 + K_6\phi_3 + K_7 a \Delta x + K_8 a \Delta y + K_9 a \Delta z + K_{10} a\phi_1 + K_{11} a\phi_2 + K_{12} a\phi_3 + K_{13} b \Delta x + K_{14} b \Delta y + K_{15} b \Delta z + K_{16} b\phi_1 + K_{17} b\phi_2 + K_{18} b\phi_3 \quad (34)$$

where the K_i are constants obtained by evaluating contour integrals around C_1 whose integrands involve F, $\partial F/\partial X$, $\partial F/\partial Y$, X, Y, dX/ds, and dY/ds and where $c = 1$ has been set to fix the global scale factor. The detailed formulas for K_i are given in Kanatani (1985). Equation 34 is nonlinear in the eight unknowns: Δx, Δy, Δz, ϕ_1, ϕ_2, ϕ_3, a, and b. To find these unknowns, eight or more different functions $F(X, Y)$ are first chosen. For each function one can calculate ΔI and the K_i to get one Eq. 34. Then one finds the least-square solution of the set of eight or more equations 34. Whether a unique solution can be obtained by this method is yet to be answered.

Orthographic Projections. For orthographic projections, instead of Eq. 3, one has

$$X = x \qquad Y = y \quad (35)$$

Again, assume that the points observed lie on a plane in 3-D whose equation at t_1 is

$$ax + by + cz = 1 \quad (13)$$

Then, from Eqs. 2, 35, and 13 (Young and Wang, 1984),

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = A \begin{bmatrix} X \\ Y \end{bmatrix} + D$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \triangleq \begin{bmatrix} r_{11} - ar_{13} & r_{12} - br_{13} \\ r_{21} - ar_{23} & r_{22} - br_{23} \end{bmatrix}$$

and

$$D = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \triangleq \begin{bmatrix} r_{13} + \Delta x \\ r_{23} + \Delta y \end{bmatrix}$$

($c = 1$ has been set to fix the global scale factor). Thus, the relationship between (X, Y) and (X', Y') is an affine transformation. This should be contrasted with the case of central projections where the relationship is Eq. 16.

One can attempt to find the motion and structure parameters (n_1, n_2, θ, Δx, Δy, a, and b) in two steps: first, from a contour correspondence over two frames, determine A and D in the affine transform Eq. 36 (a contour correspondence implies no point correspondences between the contour pair) and, second, determine the desired parameters from A and D. Several techniques for carrying out step 1 have been proposed. Reference 34 describe a method that relates the moment tensors of the two regions bounded by the contours at t_1 and t_2, respectively; Cyganski and Orr (1985) describes a method that relates the Fourier coefficients of the two contours after a canonic parameterization. A related work is Kanatani (1985). Unfortunately, step 2 is generally not possible. The unknown parameters cannot be determined from A and D without additional information. This is because there are six equations:

$$r_{11} - ar_{13} = a_{11} \qquad r_{12} - br_{13} = a_{12}$$
$$r_{21} - ar_{23} = a_{21} \qquad r_{22} - br_{23} = a_{22}$$
$$r_{13} + \Delta x = d_1 \qquad r_{23} + \Delta y = d_2$$

but seven unknowns: n_1, n_2, θ, Δx, Δy, a, and b. Solution becomes possible if one is given, eg, (a, b), ie, the orientation of the plane at t_1.

To close this section, note the classical result of Ullman (1979) for the orthographic projection case: four-point correspondence over three views determine motion/structure uniquely.

MOTION FROM OPTICAL FLOW

Problem Statement

In the two-view case, if $t_2 - t_1 = \Delta t$ is small,

$$R \approx I + S \Delta t \quad (37)$$

where

$$S \triangleq \begin{bmatrix} 0 & -w_3 & w_2 \\ w_3 & 0 & -w_1 \\ -w_2 & w_1 & 0 \end{bmatrix}$$

and I is a 3×3 unity matrix.

The symbol Ω is used to denote the vector (w_1, w_2, w_3), the instantaneous angular velocities around the x, y, and z axes, respectively, at t_1. Also,

$$T \approx v \Delta t$$

where

$$v \triangleq \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix}$$

are the instantaneous translational velocities along the axes at t_1. Letting $\Delta t \to 0$, Eq. 2 becomes

$$\frac{dp(t)}{dt} = S(t)p(t) + v(t) \quad \text{(matrix equation)} \quad (38)$$

or, equivalently,

$$\frac{dp(t)}{dt} = \Omega(t) \times p(t) + v(t) \quad \text{(vector equation)} \quad (39)$$

where

$$p(t) \triangleq \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} \quad (40)$$

As $\Delta t \rightarrow 0$, in the image plane.

$$V_x \triangleq \lim_{\Delta t \to 0} \frac{\Delta X}{\Delta t} = \frac{dX}{dt} \qquad V_y \triangleq \lim_{\Delta t \to 0} \frac{\Delta Y}{\Delta t} = \frac{dY}{dt} \tag{41}$$

The image-plane velocity vector (V_x, V_y) is referred to as the optical flow. The problem of interest is to determine v (to within a scale factor) and Ω at time t_1 from optical-flow information. One takes an approach similar to that of using point correspondences in the two-view case. Specifically, it consists of two steps: Find optical-flow vectors at N image points, $[(X_i, Y_i), (Vx_i, Vy_i)]$, $i = 1, 2, \ldots, N$ and solve equations obtained from the optical-flow information to determine v and Ω.

Finding Optical Flow

Two approaches to finding optical flow are described. The first approach is to find point correspondences between two image frames at t_1 and t_2 (with $t_2 - t_1 = \Delta t$ small) using methods discussed above, and then obtain the optical-flow vectors by

$$V_x \approx \frac{\Delta X}{\Delta t} \; V_y \approx \frac{\Delta Y}{\Delta t} \tag{42}$$

The second approach is to relate temporal and spatial differences of the image brightness. Let $f_1(X, Y)$ and $f_2(X, Y)$ be the brightness at point (X, Y) in the two successive image frames (at t_1 and t_2, respectively). At any given image point (X_0, Y_0) the time (frame) difference is

$$\Delta f(X_0, Y_0) \triangleq f_2(X_0, Y_0) - f_1(X_0, Y_0) \tag{43}$$

Assume the image point (X_0, Y_0) at t_1 and the image point (X_0', Y_0'), at t_2 correspond to the same physical point on the 3-D object, and let

$$\Delta X = X_0' - X_0 \qquad \Delta Y = Y_0' - Y_0$$

Then

$$\Delta f(X_0, Y_0) = f_2(X_0, Y_0) - f_2(X_0', Y_0') \tag{44}$$

if one makes the assumption that any given point on the 3-D object appears in the two image frames with the same brightness. If the motion is small, this brightness-constancy assumption is reasonable in many situations. Then $f_2(X_0', Y_0') = f_2(X_0 + \Delta X, Y_0 + \Delta Y)$ is expanded into a Taylor series around (X_0, Y_0) and only the linear terms are kept to get

$$\Delta f(X_0, Y_0) = -\Delta X \frac{\partial f}{\partial X}(X_0, Y_0) - \Delta Y \frac{\partial f}{\partial Y}(X_0, Y_0) \tag{45}$$

This is an important equation mentioned again in the section Motion Estimation by Direct Matching of Image Intensities. Here, one can use it to find optical flow in the following way (Rocca, 1972; Limb and Murphy, 1975). If there are two or more image points (near each other) that one can assume to have the same $(\Delta X, \Delta Y)$, by calculating Δf and $[\partial f/\partial X, \partial f/\partial Y]$ (using a difference approximation) at each point, one can get a set of linear equations in the two unknowns ΔX and ΔY. Finally, the least-squares solution of these linear equations is found, and Eq. 42 is used to get V_x and V_y.

For general 3-D motion $(\Delta X, \Delta Y)$ vary with (X, Y). Therefore, it may not be reasonable to assume that $(\Delta X, \Delta Y)$ are the same at several image points. Horn and Schunck (1981) considered the case where $(\Delta X, \Delta Y)$ change slowly with (X, Y) and formulated a variational method for estimating $(\Delta X, \Delta Y)$. Other methods that are image-point-wise recursive are described (Robbins and Netravali, 1983; Cafforio and Rocca, 1983). Also, Nagel (1983) attempted to improve the estimation of $(\Delta X, \Delta Y)$ by including the second-order terms in the Taylor series expansion of $f_2(X_0 + \Delta X, Y + \Delta Y)$. For a recent insightful study on the determination of optical flow, see Hildreth (Hildreth, 1984). See also the pioneering work on optical flow by Prazdny (1980).

Basic Equations

Differentiating Eq. 3 with respect to t and using Eq. 38, one gets

$$V_x = \left\{\frac{V_x}{z} - X\frac{V_z}{z}\right\} + [-XYw_1 + (1 + X^2)w_2 - Yw_3] \tag{46}$$

$$V_y = \left\{\frac{V_y}{z} - Y\frac{V_z}{z}\right\} + [-(1 + Y^2)w_1 - XYw_2 + Xw_3]$$

whence

$$z = \frac{v_x - Xv_z}{V_x + XYw_1 - (1 + X^2)w_2 + Yw_3} = \frac{v_y - Yv_z}{V_y + (1 + Y^2)w_1 - XYw_2 + Xw_3} \tag{47}$$

and

$$(v_x - Xv_z)[V_y + (1 + Y^2)w_1 - XYw_2 - Xw_3] = (v_y - Yu_z)[V_x + XYw_1 - (1 + X^2)w_2 + Yw_3] \tag{48}$$

Equation 48 is nonlinear in the six unknowns: v_x, v_y, v_z, w_1, w_2 and w_3. Also, it is homogeneous in v_x, v_y, and v_z. Therefore, $v = (v_x, v_y, v_z)$ can be determined only to within a scale factor. To fix ideas, let the sought after translation be the unit translation vector

$$\hat{v} = (\hat{v}_x, \hat{v}_y, \hat{v}_z) \triangleq \frac{v}{\sqrt{v_x^2 + v_y^2 + v_z^2}} \tag{49}$$

Then Eq. 48 contains five unknowns, eg, $\hat{v}_x, \hat{v}_y, w_1, w_2, w_3$. If there are optical-flow vectors at five or more image points, $[(X_i, Y_i), (Vx_i, Vy_i)]$, $i = 2, \ldots, N$, one can seek a least-squares solution to the set of N nonlinear equations 48. Note that Eq. 46 can be derived from Eq. 4 by letting $\Delta t \rightarrow 0$.

A Linear Algorithm

Similar to the two-view point-correspondence case, a linear algorithm is possible here (Zhuang and co-workers, in

press). In fact, in Eq. 12, if one sets

$$G = K\,\Delta t \tag{50}$$

where

$$K \triangleq \begin{bmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{bmatrix} \tag{51}$$

and

$$R = \mathrm{I} + S\,\Delta t, \tag{37}$$

and then lets $\Delta t \to 0$, one gets

$$[V_X, V_Y, 0]K \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} + [X, Y, 1]KS \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{52}$$

Let

$$L = \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \triangleq KS \tag{53}$$

Then Eq. 52 is equivalent to

$$[X^2, Y^2, 1, XY, X, Y, V_y, -V_x, V_xY - V_yX]h = 0 \tag{54}$$

where

$$h = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \end{bmatrix} \triangleq \begin{bmatrix} l_{11} \\ l_{22} \\ l_{33} \\ l_{12} + l_{21} \\ l_{13} + l_{31} \\ l_{23} + l_{32} \\ v_x \\ v_y \\ v_z \end{bmatrix} \tag{55}$$

From Eqs. 53 and 55

$$\begin{bmatrix} h_7 \\ h_8 \\ h_9 \end{bmatrix} = \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \tag{56}$$

and

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \end{bmatrix} = \begin{bmatrix} 0 & -w_2 & -w_3 \\ -w_1 & 0 & -w_3 \\ -w_1 & -w_2 & 0 \\ w_2 & w_1 & 0 \\ w_3 & 0 & w_1 \\ 0 & w_3 & w_2 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \tag{57}$$

The solution procedure is as follows. From eight or more optical-flow vectors, one determines $h_1, \ldots, h_9$ to within a scale factor from the linear Eq. 54. Then Eq. 56 gives $v = (v_x, v_y, v_z)$ to within a scale factor. Finally, Eq. 57 is used to find $\Omega = (w_1, w_2, w_3)$.

Planar Patch Case. The linear algorithm of the last section breaks down when all the image points under consideration correspond to 3-D points lying on a plane (Longuet-Higgins, 1984). However, similar to the two-view case, a different linear algorithm is available. Let the equation of the plane in 3-D be

$$ax + by + cz = 1 \tag{13}$$

Then, as before,

$$\frac{1}{z} = aX + bY + c \tag{17}$$

Substituting in Eq. 46, one gets

$$\begin{aligned} V_X &= k_1 + k_2X + k_3Y + k_7X^2 + k_8XY \\ V_Y &= k_4 + k_5X + k_6Y + k_7XY + k_8Y^2 \end{aligned} \tag{58}$$

where

$$\begin{aligned} &k_1 = cv_x + w_2 \quad k_2 = av_x - cv_z \quad k_3 = bv_x - w_3, \\ &k_4 = cv_y - w_1 \quad k_5 = av_y + w_3 \quad k_6 = bv_y - cv_z, \\ &k_7 = -av_2 + w_2 \quad k_8 = -bv_z - w_3 \end{aligned} \tag{59}$$

Given optical-flow vectors at four or more image points, we can determine $k_1, k_2, \ldots, k_8$ from Eq. 58. Then $\hat{v}$ *and* Ω can be found from the k_i as described in Longuet-Higgins (1984). Similar to the two-view case, generally there are two solutions for the motion parameters. Longuet-Higgins (1984) discusses the physical meaning of the two solutions and the fact that in many cases one of the solutions can be ruled out.

Generalized Flow Fields

Basic Equations. In the discussions of optical flow so far, only the image-point velocities V_X and V_Y have been used. A more general formulation using V_X, V_Y as well as their derivatives (with respect to X and Y) up to the second order was proposed by Waxman and Ullman. Their approach is based on studying the deformation of a small neighborhood in the image and provides much insight into the relationship between the 3-D motion/structure of a rigid body and its 2-D perspective views.

Specifically, consider the vicinity of the image origin $(X, Y) = (0, 0)$, and assume that the object surface

$$z = z(x, y) \tag{60}$$

around the point $(0, 0, z_0)$, where $z_0 = z(0, 0)$ is smooth (twice differentiable). Then 12 observables can be defined that are expressible in terms of the six motion parameters $(M_1$–$M_6)$,

$$\frac{v_x}{z_0}, \frac{v_y}{z_0}, \frac{v_z}{z_0}, w_1, w_3, w_2$$

and five structure parameters $(T_1$–$T_5)$,

$$\left[\frac{\partial z}{\partial x}\right]_0, \left[\frac{\partial z}{\partial y}\right]_0, z_0\left[\frac{\partial^2 z}{\partial x^2}\right]_0, z_0\left[\frac{\partial^2 z}{\partial y^2}\right]_0, z_0\left[\frac{\partial^2 z}{\partial x\,\partial y}\right]_0$$

The subscript 0 indicates that the derivative is evaluated at $(0, 0, z_0)$. Note that the five structure parameters give information on the slopes and the curvatures of the surface at $(0, 0, z_0)$.

The 12 observables are $(0_1$–$0_{12})$ V_X, V_Y, e_{11}, e_{22}, e_{12}, w, $\partial e_{11}/\partial X$, $\partial e_{11}/\partial Y$, $\partial e_{22}/\partial X$, $\partial e_{22}/\partial Y$, $\partial w/\partial X$, and $\partial w/\partial Y$, where e_{11}, e_{22}, e_{12}, and w are defined as follows: Let

$$\frac{\partial V_i}{\partial \xi_j} = \frac{1}{2}\left[\frac{\partial V_i}{\partial \xi_j} + \frac{\partial V_j}{\partial \xi_i}\right] + \frac{1}{2}\left[\frac{\partial V_i}{\partial \xi_j} - \frac{\partial V_j}{\partial \xi_i}\right] \triangleq e_{ij} + w_{ij}$$

$$i, j = 1, 2; (V_1, V_2) = (V_X, V_Y); (\xi_1, \xi_2) = (X, Y) \quad (61)$$

In terms of image deformation, e_{ij} is the rate-of-strain tensor and w_{ij} the spin tensor. The physical meaning of these quantities are e_{11} is the rate of stretch of a differential image line oriented along the X axis, e_{22} the rate of stretch of a differential image line oriented along the Y axis, e_{12} $(= e_{21})$ one-half the rate of decrease of the angle between two differential line segments along the image axes, and w_{21} $(= -w_{12} = w)$ the rate of rotation (ie, the spin) of the differential neighborhood of image about the origin.

The basic flow equations relating the observables to the motion and structure parameters are derived in Waxman and Ullman (1983).

$$\begin{aligned}
0_1 &= M_1 + M_5 \\
0_2 &= M_2 - M_4 \\
0_3 &= -M_3 - M_1T_1 \\
0_4 &= -M_3 - M_2T_2 \\
0_5 &= -\tfrac{1}{2}(M_2T_1 + M_1T_2) \\
0_6 &= M_6 + \tfrac{1}{2}(M_1T_2 - M_2T_1) \\
0_7 &= 2(M_5 + M_3T_1) - M_1T_3 \\
0_8 &= -M_4 + M_3T_2 - M_1T_5 \\
0_9 &= M_5 + M_3T_1 - M_2T_5 \\
0_{10} &= 2(M_3T_2 - M_4) - M_2T_4 \\
0_{11} &= \tfrac{1}{2}(M_4 - M_3T_2 - M_2T_3 + M_1T_5) \\
0_{12} &= \tfrac{1}{2}(M_5 + M_3T_1 + M_1T_4 - M_2T_5)
\end{aligned} \quad (62)$$

These flow equations form a set of 12 coupled nonlinear algebraic equations with 11 unknowns. A method of solving these equations (given 0_1–0_{12}) is described in Waxman and Ullman (1983).

Finding the Observable. The problem remains: How does one measure the observables from the image sequence? Kanatani (1985), Waxman and Wohn (1984a,b) suggest a method based on evolving contours in the image plane. The 12 observables are in terms of $V_X^{(i,j)}$ and $V_Y^{(i,j)}$, i, $j = 0, 1, 2$ and $i - j \leq 2$, where

$$V_X^{(i,j)} \triangleq \left.\frac{\partial^{i+j} V_X}{\partial X^i\,\partial Y^j}\right|_0 \quad (63)$$

and similarly for $V_Y^{(i,j)}$. These derivatives can be obtained in the following manner. In the vicinity of $(X, Y) = (0, 0)$, one can write

$$\begin{aligned}
V_X(X, Y) &= \sum_{\substack{i=0 \\ (i+j\leq 2)}}^{2}\sum_{j=0}^{2} V_X^{(i,j)}\frac{X^i}{i!}\frac{Y^j}{j!} \\
V_Y(X, Y) &= \sum_{\substack{i=0 \\ (i+j\leq 2)}}^{2}\sum_{j=0}^{2} V_Y^{(i,j)}\frac{X^i}{i!}\frac{Y^j}{j!}
\end{aligned} \quad (64)$$

For curved surfaces Eqs. 64 are only locally (and approximately) valid. But for planes they are globally valid—see Eqs. 58.

Assume a planar contour is tracked over two image frames separated by a small Δt. If one measures at a point (X, Y) on the contour, the normal flow velocity $V_n(X, Y)$ and the normal of the contour $n(X, Y) = (n_X, n_Y)$, one gets the equation

$$V_n(X, Y) = \sum_{\substack{i=0 \\ (i+j\leq 2)}}^{2}\sum_{j=0}^{2} \frac{X^i}{i!}\frac{Y^j}{j!}\{n_X(X, Y)V_X^{(i,j)} + n_Y(X, Y)V_Y^{(i,j)}\} \quad (65)$$

Since there are 12 unknowns, one needs to measure the V_n and n of at least 12 points on the contour. Note that several separate contours can be used as long as they lie in the same plane in 3-D.

For curved surfaces the problem is much more difficult. Waxman and Wohn (1984b) discusses the truncation errors incurred by using the approximate Eqs. 64.

MOTION ESTIMATION BY DIRECT MATCHING OF IMAGE INTENSITIES

All the techniques for 3-D motion determination described above fall into the category of two-step methods. First, correspondences or optical-flow vectors are found, and then equations are solved to obtain the motion/structure parameters. In this section a description of a method based on direct matching of image intensities is given (also, see Finding Optical Flow, above).

Determining 2-D Translation by Displaced Frame Differences

Consider first the simple case of 2-D translation, ie, assume that $(\Delta X, \Delta Y)$ is constant for all image points corresponding to physical points on the rigid body. Again, let

$f_1(X, Y)$ = brightness of first frame (at t_1)

$f_2(X, Y)$ = brightness of second frame (at t_2)

Then the approach is to match f_1 and f_2 directly: Find $(\Delta X, \Delta Y)$ to minimize $D\{f_1(X, Y), f_2(X - \Delta X, Y + \Delta Y)\}$, where

D is a distance measure. One commonly used distance measure is

$$D = \sum_{X,Y} \{f_1(X, Y) - f_2(X + \Delta X, Y + \Delta Y]\}^2 \quad (66)$$

It is important to point out that this direct matching approach makes the tacit assumption that the two image points at t_1 and t_2, respectively, corresponding to the same physical point on the object, have the same brightness: ie, the brightness of an image point corresponding to a fixed point on the object does not change after motion. This is called the brightness-constancy assumption.

Coming back to Eq. 66, one notes that D can be minimized by using standard optimization techniques. However, the computation can be simplified in the case where the motion $(\Delta X, \Delta Y)$ is small. Then one can expand $f_2(X + \Delta X, Y + \Delta Y)$ in a Taylor series around (X, Y) and retain up to only the first-order terms. And Eq. 66 is reduced to

$$D = \sum_{X,Y} \left(\Delta f + \Delta X \frac{\partial f_2}{\partial X} + \Delta Y \frac{\partial f_2}{\partial Y}\right)^2 \quad (67)$$

where

$$\Delta f(X, Y) \triangleq f_2(X, Y) - f_1(X, Y)$$

is the frame difference at (X, Y) (Robbins and Netravali, 1983; Caffario and Rocca, 1983).

In practice, Δf and $\partial f_2/\partial X$, $\partial f_2/\partial Y$ is calculated at N points: (X_i, Y_i), $i = 1, 2, \ldots, N$. Then the summation in Eq. 67 will be over these N points. Note that minimizing D in Eq. 67 is equivalent to finding the least-squares solution of the set of linear equations:

$$-(\Delta f)_i = \Delta X \left(\frac{\partial f_2}{\partial X}\right)_i + \Delta Y \left(\frac{\partial f_2}{\partial Y}\right)_i \qquad (i = 1, 2 \ldots, N) \quad (68)$$

where a subscript i indicates that the quantity is evaluated at (X_i, Y_i). This is the same as the method described in Finding Optical Flow.

Generalization to 3-D Motion

The method of the preceding section can in principle be extended to the general case of 3-D motion. Both ΔX and ΔY are expressed in terms of the 3-D motion parameters; then D in Eq. 66 is minimized with respect to the 3-D motion parameters. In practice, there are two difficulties. The first is computational: There must be searching in a high-dimensional space. The second is that (as shown below), without further assumptions, the number of solutions is infinite. From Eqs. 1 and 4 one can get ΔX and ΔY in terms of, eg, X, Y, $z/\Delta z$, n_1, n_2, θ, $\Delta x/\Delta z$, and $\Delta y/\Delta z$ (assuming $\Delta z \neq 0$). Then D in Eq. 66 is minimized with respect to these latter variables. Unfortunately, for each point (X_i, Y_i) there is a new unknown $z_i/\Delta z$. Therefore, one always has five more unknowns (the motion parameters) than the number of terms in Eq. 66, and as a result one has infinitely many solutions to the minimization problem.

One can hope to get a unique solution if one knows the form of the object surface to within a finite number of parameters. The simplest case is when the surface is a plane. Then it can be represented by

$$ax + by + cz = 1 \quad (\text{at } t_1) \qquad (13)$$

and

$$\frac{z}{\Delta z} = \frac{1}{a'X + b'Y + c'} \quad (69)$$

where

$$a' \triangleq a\,\Delta z, \qquad b' \triangleq b\,\Delta z \qquad c' \triangleq c\,\Delta z \quad (70)$$

As a result, D in Eq. 66 can be expressed in terms of the eight unknown parameters a', b', c', n_1, n_2, θ, $\Delta x/\Delta z$, and $\Delta y/\Delta z$ independent of how many points (X_i, Y_i) are used in the summation.

Now the computational problem: To search in an eight-dimensional space by standard optimization techniques is very time-consuming. The situation is better if the 3-D motion is small so that all $(\Delta X, \Delta Y)$ are small. Then one can use the Taylor series approach, and the problem of minimizing D is reduced to the problem of finding the least-squares solution of the set of Eq. 68, where ΔX and ΔY are now written in terms of the eight unknowns mentioned above. Note that the equations are now nonlinear (Huang and Tsai, 1981; Huang, 1985).

To summarize, the method of determining 3-D motion parameters of a rigid planar patch is to calculate Δf and $\delta f_2/\delta X$, $\delta f_2/\delta Y$ at eight or more ponts, and then find the least-squares solution (by some iterative method) of the set of eight or more nonlinear Eq. 68, where ΔX and ΔY are written in terms of the eight unknowns a', b', c', n_1, n_2, θ, $\Delta x/\Delta z$, and $\Delta y/\Delta z$ by using Eqs. 4 and 69.

Once again, note that the method assumes brightness constancy.

Linear Algorithm for Planar Patches

The nonlinear least-squares algorithm for determining 3-D motion parameters of a rigid planar patch as described in the preceding section can be reduced to a linear least-squares problem by introducing appropriate intermediate variables (Tsai and Huang, 1981). Specifically, from Eq. 16

$$\Delta X = X' - X = \frac{a_1X + a_2Y + a_3 - a_7X^2 - a_8XY - a_9X}{a_7X + a_8Y + a_9}$$

$$\Delta Y = Y' - Y = \frac{a_4X + a_5Y + a_6 - a_7XY - a_8Y^2 - a_9Y}{a_7X + a_8Y + a_9} \quad (71)$$

Assuming the motion to be small, one can substitute Eq. 71 into Eq. 68 to get

$$(a_7X + a_8Y + a_9)\,\Delta f = (a_1X + a_2Y + a_3 - a_7X^2 - a_8XY - a_9X)\frac{\delta f_2}{\delta X} + (a_4X + a_5Y + a_6 - a_7XY - \alpha_8Y^2 - a_9Y)\frac{\delta f_2}{\delta Y}$$

or

$$X\frac{\delta f_2}{\delta X}a_1 + Y\frac{\delta f_2}{\delta X}a_2 + \frac{\delta f_2}{\delta X}a_3 + X\frac{\delta f_2}{\delta Y}a_4 + Y\frac{\delta f_2}{\delta Y}a_5 + \frac{\delta f_2}{\delta Y}a_6 - \left[X^2\frac{\delta f_2}{\delta X} + XY\frac{\delta f_2}{\delta Y} + X\,\Delta f\right]a_7 - \left[XY\frac{\delta f_2}{\delta X} + Y^2\frac{\delta f_2}{\delta Y} + Y\,\Delta f\right]a_8 - \left[X\frac{\delta f_2}{\delta X} + Y\frac{\delta f_2}{\delta Y} + \Delta f\right]a_9 = 0 \quad (72)$$

This equation is linear and homogeneous in the nine unknowns, $a_1, \ldots, a_9$. If one calculates Δf and $\delta f_2/\delta X$, $\delta f_2/\delta Y$ at eight or more image points (X, Y), one gets a set of eight or more equations eg, Eq. 72. Then $a_1, \ldots, a_9$ can be solved to within a scale factor. Recall that the a_i are related to the motion/structure parameters by Eq. 15 and that the latter can be obtained from the former by a method described in Tsai and co-workers (1983).

MOTION FROM 3-D FEATURE CORRESPONDENCES

The motion-estimation techniques described above are based on images taken by a monocular 2-D sensor such as a single television camera. With such an arrangement the 3-D translation and the range of the object can be determined to only within a scale factor. One can determine the absolute translation velocity and ranges of object points if binocular vision (see stereo vision) is used, eg, two television cameras with known relative positions and orientations. the binocular method has several other advantages described below.

Binocular Procedure

A pair of stereo images is taken at t_1, and another pair is taken at t_2, and then the following procedure is used.

1. From the two images taken at t_1 feature points are extracted, the two point patterns are matched to find correspondences, and then by triangulation the 3-D coordinates of these points are found. The same is done for the two images taken at t_2.
2. The two 3-D point patterns at t_1 and t_2 are matched to find 3-D point correspondences.
3. A set of equations involving the motion parameters are obtained from the 3-D point correspondences. These equations are solved to determine motion (Huang and Blostein, 1985).

Note that the matching problems in 1 and 2 are usually easier than the matching problem in the monocular two-view case (see above) because in 1, for a fixed point in one image of the stereo pair, the corresponding point in the other image is restricted to lie on the so-called epipolar line, and in 2, the distances between pairs of the 3-D points on a rigid body is invariant to motion. An algorithm for the maximal matching of two 3-D point sets is presented by Chen and Huang (1986).

Motion from 3-D Correspondences

Once one has obtained 3-D point correspondences $p_i \longleftrightarrow p_i'$, $i = 1, 2, \ldots, N$, where

$$p = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad \text{and} \quad p' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}$$

how does one get the motion parameters R and T? A related question is: What is the minimum number of 3-D point correspondences needed for unique determination of R and T of a rigid body? A basic fact is that R and T are determined uniquely by three 3-D point correspondences (assuming the three points are not collinear). This becomes obvious if one notes that two points will fix a rigid body in space except for a possible rotation around the axis formed by joining the two points. A third point then fixes the rigid body completely. Once one knows three 3-D point correspondences on a rigid body, one can generate any number of other 3-D point correspondences rigid relative to the original three points.

To describe algorithms for finding R and T, Eq. 2 is rewritten as

$$p' = Rp + T \quad (73)$$

There are six unknown parameters, n_1, n_2, θ, Δx, Δy, and Δz. Each 3-D point correspondence gives one matrix Eq. 73 or three scalar equations, which are nonlinear in the unknowns. An obvious method would be to find the least-squares solution (by some iterative technique) of the set of $3N$ coupled nonlinear equations obtained from the N three-dimensional point correspondences, where $N \geq 3$. However, much simpler linear algorithms are available (Blostein and Huang, 1984), one of which is described below.

Assume there are three 3-D point correspondences

$$p_i \longleftrightarrow p_i', \qquad i = 1, 2, 3.$$

Let

$$m_1 \triangleq p_1 - p_3 \qquad m_1' \triangleq p_1' - p_3'$$
$$m_2 \triangleq p_2 - p_3 \qquad m_2' \triangleq p_2' - p_3' \quad (74)$$

Then, from Eq. 73,

$$m_1' = Rm_1 \quad m_2' = Rm_2 \quad (75)$$

If

$$m_3 \triangleq m_1 \times m_2 \qquad m_3' \triangleq m_1' \times m_2' \tag{76}$$

(Consider m_i and m_i' as vectors.), then

$$m_3' = Rm_3 \tag{77}$$

Combining Eqs. 75 and 77,

$$[m_1', m_2', m_3'] = R[m_1, m_2, m_3] \tag{78}$$

whence

$$R = [m_1', m_2', m_3'][m_1, m_2, m_3]^{-1} \tag{79}$$

and

$$T = p_i' - Rp_i \quad \text{for } i = 1, 2, 3 \tag{80}$$

Note that the numerical accuracy of this algorithm is usually improved if normalized (to a magnitude of 1) versions of m_i and m_i' are used in the formulation.

Two remarks are in order. First, the above algorithm can be used not only for 3-D point correspondences but also for 3-D straight-line correspondences and surface-normal correspondences. In the latter two cases only two correspondences are needed. Second, in the pressure of noise in the data (3-D point coordinates), the matrix R obtained from the above algorithm may not be a rotation (ie, orthonormal and with a determinant equal to +1). In that case a rotation matrix R' can be found by using the algorithms in Faugeras and Hebert 91983) and Huang and co-workers (1986) to minimize

$$\|R' - R\|^2 \triangleq \sum_{i=1}^{3} \sum_{j=1}^{3} (r_{ij}' - r_{ij})^2$$

where r_{ij} and r_{ij}' are elements of R and R' respectively.

Correspondenceless Approaches

The problem of retinal correspondence is ill-defined and only partial solutions have been obtained to date. This complicates three-dimensional analysis on the basis of visual motion. Consider a set of points $A = \{(X_i, Y_i, Z_i), i = 1, \ldots, n\}$ in three dimensions that moves rigidly to a new position $A' = \{(X_i', Y_i', Z_i'), i = 1, \ldots, n\}$. Given the images $A_I = \{(x_i, y_i), i = 1, \ldots, n\}$, $A_I' = \{(x_i', y_i'), i = 1, \ldots, n\}$ (before and after the motion) and without considering individual point correspondences (only correspondences of sets of points), the problem is to recover the 3-D motion involved. Various approaches can be found (Aloimonos, 1986; Aloimonos and Rigoutsos, 1986; Aloimonos and Hervé, 1990).

Motion and Shape from Normal Flow

Although the idea of the normal optical flow field has existed in the literature for quite some time, using it to extract information about 3-D motion and structure is a rather recent activity. This is a very promising research area since normal flow fields are much easier to compute than actual flow fields. It turns out that if one employs an active observer (an observer that can control the geometric parameters of its sensory apparatus) then 3D motion and structure can be computed from normal flow. The interested reader can consult (Aloimonos, Weiss, and Bandyopadhyay (1987) and Aloimonos (1990).

ADDITIONAL TOPICS

In the preceding sections the major approaches to determining 3-D motoin/structure of a rigid body are described in some detail. This last section is a brief commend on some important additional topics. These topics also represent areas where further research is needed.

Numerical Accuracy of Algorithms

The reader should be warned that computer simulations and experiments with real images (Fang and Huang, 1984a,b) have indicated that in order to estimate motion parameters reasonably accurately (around 10% error) from two perspective views using a single camera, the image resolution has to be quite high (typically 1000 × 1000 picture elements, assuming image-point features can be measured to within one picture element). Theoretical studies or even systematic simulation studies on how the estimation errors depend on various factors are yet to be made. The situation with the two-camera case is somewhat better (Huang and Blostein, 1985). Some simulation results for the two-camera case are given below to indicate how redundant point correspondences can be used to improve estimation accuracy.

The algorithm of Motion form 3-D Correspondences (above) requires only three 3-D point correspondences. If more than three point correspondences are available, the redundancy can be used to improve estimation accuracy in several ways, two of which are adaptive least-squares (Huang and Blostein, 1985) and RANSAC (Fischler and Bolles, 1981). A hybrid of the two was used in Huang and Blostein (1985), from which some computer simulation results are quoted. The imaging geometry is as follows: Two pinhole cameras with focal length 28-mm are used, and the two image planes are coplanar; each image is 38 mm × 50 mm and has a resolution of 512 × 512 picture elements. The baseline distance between the two cameras is 400 mm.

The 3-D points are chosen randomly in a cube centered at a point 3 m from the cameras each side of which is 0.75 m long. The true motion is a rotation of 35° about an axis through the origin with direction 0.9, 0.3, and 0.316 followed by a translation of 0.8, 0.2, and 0.6 m. The simulation is done as follows. The 3-D points before and after the motion are projected onto the two images. The image coordinates of these points are quantized (with a resolution of 512 × 512). The quantized image points are then used in the method described in Motion from 3-D Feature Correspondences to estimate R and T. That is, triangulation is done using these quantized image points to obtain the 3-D coordinates of the points, which are then used in the algo-

rithm described above. The errors in the estimated R and T are due to the inaccuracies in the 3-D coordinates of the points, which are in turn due to the quantization of the image coordinates. The results are: The average errors (in %) of θ, n_1, n_2, n_3, Δx, Δy, and Δz are, respectively: 5.2, 2.3, 14.5, 8.1, 10.1, 30.7, and 10.7 with seven 3-D point correspondences and 2.2, 1.0, 7.1, 3.1, 4.8, 14.9, and 4.4 with fifteen 3-D point correspondences. For each of the two cases the averages are computed over 100 trials.

At this point it is worth noting recent research (Spetsakis & Aloimonos, 1989) that develops algorithms that are provably optimal under assumptions about the noise that corrupts retinal correspondences. Experiments with such algorithms demonstrate that a 1% error in the input (about 4–7 pixels) can produce a 100% error in the output (3-D motion). Such results indicate that the problem of recovering 3-D motion (rotation and translation) from point correspondences in two frames might be inherently unstable.

Multiple Objects

The methods described in the earlier sections are for a single isolated rigid body. What if the scene contains several rigid bodies moving differently (this includes the special case of a single rigid body moving against a stationary but textured background)? Segmentation needs to be done somewhere along the way. If one is working with the two-view case described in Solution Using Point Correspondences (above) and if the motions of the rigid bodies are small from t_1 to t_2, the following approach can be tried.

Assuming the motions are small, one can still hope to get correct point correspondences. However, one does not know which points lie on which objects. This one attempts to find by a clustering technique. The basic ideas is to take all possible octets from the point correspondences, and for each octet compute R and $\hat{T}$ using the algorithm described above under A Linear Algorithm. Then clusters are found in the five-dimensional $(n_1, n_2, \theta, \Delta x, \Delta \hat{y})$-space. Ideally, each rigid body will give one cluster. To save computation, one uses heuristics (qv) to reduce the number of octets to consider and perhaps does clustering in subspaces of the five-dimensional space. Obviously, the same approach can be used in the binocular case. here, one only has to deal with triplets.

In order to handle the multiple-object case effectively, constraints on the scenario should be used wherever possible. A very impressive piece of work in that direction has been done by Adiv (1985).

Multiple Frames

Most of the approaches described up to now consider two or three dynamic frames. If several frames are used, it turns out that precision is greatly increased (due to redundancy). See Spetsakis (1989) for a survey and the treatment of the problem in its most general form.

Nonrigid Objects

Two cases are of particular interest: an articulated object (ie, an object comprising several rigid parts connected through various joints) and an elastic object. Some aspects of motion analysis of articulated objects have been studied by Asada, Yachida, and Tsuji (1984); O'Rourke and Badler (1980); and Webb and Aggarwal (1983a). In particular, Webb and Aggarwal investigated the case where the rotation axis can be assumed fixed in direction throughout the observed image sequence. The same authors (1983b) have also studied a special case of elastic objects where the object is assumed to be locally rigid, which implies an affine transformation between two image planes under local parallel projection. This approach is being extended by Chen (1985) to handle general elastic bodies. Finally Koenderink and VanDoorn (1985) are investigating the special case of bending deformation. The class of bending deformations encompasses all deformations that conserve distances along the surface but not necessarily through space.

THE ROLE OF THE VISUAL FIELD

During the development of the field of visual motion analysis it was observed that results were more accurate for wide (as opposed to narrow) visual fields. This observation led to the development of techniques for finding motion parameters from spherical flow fields. In Nelson and Aloimonos (1988), a theory is developed for determining the motion of an observer given the flow field over a full 360 degree image sphere. The method is based on the fact that the foci of expansion and contraction for an observer moving without rotation are 180 degrees opposed; and on the observation that if the flow field on the sphere is considered around three equators defining the three principal axes of rotation, then the effects of the three rotational motions decouple. The three rotational parameters can thus be determined independently by searching, in each case, for a rotational value for which the derotated equatorial flow field can be partitioned into disjoint 180 degree arcs of clockwise and counterclockwise flow. The direction of translation is obtained as a by-product of this analysis. Since this search is two dimensional in the motion parameters, it can be performed relatively efficiently. Because information is correlated over large distances, the method can be considered a pattern recognition rather than a numerical algorithm. The algorithm was shown to be robust and relatively insensitive to noise and to missing data. Both theoretical and empirical studies of the error sensitivity were presented. The theoretical analysis showed that for white noise of bounded magnitude M, the expected error is at worst linearly proportional to M. Empirical tests demonstrated negligible error for perturbations of up to 20% in the input, and errors of less than 20% for perturbations of up to 200%.

Motion Modeling and Prediction

This article has been concerned mainly with estimating the motion parameters R and T of an object between two time instants t_1 and t_2 based on image frames taken at these time instants. In most practical problems one is more interested in predicting rather than just estimating motion.

In order to predict, one needs a model of the motion that is valid over a number of image frames and contains a small number of parameters that remain constant over these frames. One can first estimate these parameters based on the first few frames and then use these estimated values to predict future motion and hence where the object will be in future frames.

One such approach is described in Huang, Weng, and Ahuja (1986), where the object has a precessional motion around its center of gravity, which is moving on a polynomial curve (eg, a parabola) in space.

High-Level Motion Understanding

In many cases the ultimate goal of motion analysis is to come up with a symbolic description of the dynamic scene under study. A complete system can conveniently be thought of as comprising two modules. The first module extracts from the observed raw data (eg, an image sequence), low/intermediate-level features such as motion and structure parameters. Then the second module arrives at a symbolic description of the dynamic scene by high-level reasoning based on the low/intermediate features as well as other a priori information about the scene.

One can find such complete dynamic scene-analysis systems in the literature in the biomedical area. Two excellent examples are Levine and co-workers, 1983, which describes a rule-based system for characterizing blood cell motion, and Tsotsos and co-workers, 1980, which describes a system for analyzing the motion of left-ventricle walls. In both cases the "scenes" are basically 2-D in nature, and therefore the task of the low/intermediate-level module is greatly simplified.

For truly 3-D scenes a complete dynamic scene-analysis system is hard to construct. The main problem is that the low/intermediate-level features the high-level module needs for its reasoning may be very difficult, if not impossible, to extract from the raw data. In fact, the low/intermediate-level module will probably need help from high-level reasoning to improve its performance. Some impressive examples of high-level modules are O'Rourke and Badler (1980), Neumann (1984), and Borchardt (1984). Neumann (1984) describes a system that observes traffic scenes and produces natural-language descriptions of them. In particular, the system will recognize and verbalize interesting occurrences (events) in the scene—eg, one car is overtaking another. Borchardt (1984) describes an expert system for event identification. The applications considered are simple assembly-line tasks. However, in both systems the low/intermediate-level features needed by the high-level modules are furnished at least in part by human operators.

Future Research

To summarize, the following are important research topics in motion analysis:

To find robust algorithms for motion estimation,

To find algorithms for estimating motion of multiple objects,

To find algorithms for estimating motion of nonrigid objects,

To find algorithms for predicting motion, and

To link and coordinate low/intermediate-level and high-level motion analysis.

BIBLIOGRAPHY

G. Adiv, "Determining 3-D Motion and Structure from Optical Flow Generated by Several Moving Objects," *IEEE Trans. PAMI* **7**(4), 384–410 (July 1985).

J. (Y.) Aloimonos, *Computing Intrinsic Images,* Ph.D. dissertation, University of Rochester, Rochester, N.Y., 1986.

J. (Y.) Aloimonos, "Purposive and Qualitative Active Vision," *Proceedings of the DARPA Image Understanding Workshop,* Pittsburgh, Pa., September 1990, pp. 816–828.

J. (Y.) Aloimonos and J. Y. Hervé, "Correspondenceless Stereo and Motion: Planar Surfaces," *IEEE Trans. PAMI* **12,** 504–510 (1990).

J. (Y.) Aloimonos and I. Rigoutsos, "Determining the 3-D Motion of a Rigid Surface Patch Without Correspondence Under Perspective Projection," *Proceedings of the Fifth National Conference on AI,* AAAI Menlo Park, Calif., 1986, pp. 681–686.

J. (Y.) Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active Vision," *Intl. J. Computer Vision* **1,** 333–356 (1988).

M. Asada, M. Yachida, and S. Tsuji, "Understanding of 3-D Motions in Blocks World," *Patt. Recog.* **17**(1), 57–84 (1984).

D. H. Ballard and C. M. Brown, *Computer Vision,* Prentice-Hall, Englewood Cliffs, N.J., 1982.

A. Bandopadhay, *A Computational Study of Rigid Motion Perception,* Ph.D. dissertation, University of Rochester, Rochester, N.Y., 1986.

A Bandopadhay and J. Aloimonos, "Image Motion Estimation by Clustering," *Int'l Journal of Imaging Science and Technology* (in press).

S. D. Blostein and T. S. Huang, "Estimating Motion from Range Data," *Proceedings of the First Conference on AI Applications,* Denver, Colo., 1984.

G. C. Borchardt, *A Computer Model for the Representation and Identification of Phyical Events.* Technical Report T-142, Coordinated Science Laboratory, University of Illinois, Urbana, May 1984.

C. Cafforio and F. Rocca, "The Differential Method for Image Motion Estimation," in Huang, 1983.

S. S. Chen, "Shape and Correspondence of Nonrigid Objects," *Proceedings of the IEEE Workshop on Computer Vision,* Bellaire, Mich., 1985.

J. K. Cheng and T. S. Huang, "Image Registration by Matching Relational Structures," *Patt. Recog.* **17**(1), 149–160 (1984).

H. H. Chen and T. S. Huang, "Maximal Matching of Two 3-D Point Sets," *Proceedings of the International Conference on Pattern Recognition,* Paris, 1986.

Comput. Vis. Graph. Img. Proc., Special issues on Motion and Time-Varying Imagery **21**(1 and 2), 1–293 (Jan. and Feb. 1983).

D. Cyganski and J. A. Orr, "3-D Motion Parameters from Contours Using a Canonic Differential," *Proceedings of the ICASSP 85,* 1985, pp. 24.91–4.

R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis,* Wiley, New York, 1973, p. 373.

J. Q. Fang and T. S. Huang, *A Corner-Finding Algorithm for*

Image Analysis and Registration, Pittsburgh, Pa., August 18–20, 1982, pp. 46–49.

J. Q. Fang and T. S. Huang, "Some Experiments on Estimating the 3-D Motion Parameters of a Rigid Body From Two Consecutive Image Frames," *IEEE Trans. PAMI* **6**(5), 547–554 (Sept. 1984).

J. Q. Fang and T. S. Huang, "Solving 3-D Small-Rotation Motion Equations," *Comput. Vis. Graph. Img. Proc.* **26,** 183–296 (1984a).

J. Q. Fang and T. S. Huang, "Some Experiments on Estimating the 3-D Motion Parameters of a Rigid Body From Two Consecutive Image Frames," *IEEE Trans. PAMI* **6**(5), 547–555 (Sept. 1984b).

O. D. Faugeras and M. Hebert, "A 3-D Recognition and Positioning Algorithm Using Geometrical Matching between Primitive Surfaces," *Proceedings of the Eighth IJCAI,* Morgan-Kaufmann, San Mateo, Calif., 1983, pp. 996–1102.

M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography," *CACM* **24**(6), 381–395 (June 1981).

J. P. Gambotto and T. S. Huang, "Motion Analysis of Isolated Targets in Infrared Image Sequences," *Proceedings of the Seventh ICPR,* Montreal, 1984.

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

W. K. Gu, J. Y. Yang, and T. S. Huang, "Matching Perspective Views of a 3-D Object Using Composite Circuits," *Proceedings of the Seventh ICPR,* 1984.

E. C. Hildreth, *The Measurement of Visual Motion,* MIT Press, Cambridge, Mass., 1984.

B. Horn, *Robot Vision,* McGraw Hill, New York, 1986.

B. K. P. Horn and B. G. Schunck, "Determining Optical Flow," *Artif. Intell.* **17,** 185–203 (1981).

T. S. Huang, ed., *Image Sequence Analysis,* Springer-Verlag, Heidelberg, Germany, 1981.

T. S. Huang, ed., *Image Sequence Processing and Dynamic Scene Analysis,* Springer-Verlag, Heidelberg, Germany, 1983.

T. S. Huang, "Determining 3-D Motion/Structure from Two Perspective Views," in T. Y. Young and K. S. Fu, eds., *Handbook of Pattern Recognition and Image Processing,* Academic Press, New York, 1985.

T. S. Huang and R. Y. Tsai, "Image Sequence Analysis: Motion Estimation," in Huang, 1981.

T. S. Huang, "Three-Dimensional Motion Analysis by Direct Matching," *Conference Digest, Optical Society of America Topical Meeting on Computer Vision,* Incline Village, Nevada, 1985, pp. FAI-1–4.

T. S. Huang and S. D. Blostein, "Robust Algorithms for Motion Estimation Based on Two Sequential Stereo Image Pairs," *Proceedings of the Conference on Computer Vision and Pattern Recognition,* San Francisco, 1985.

T. S. Huang, J. Weng, and N. Ahuja, "3-D Motion from Image Sequences: Modeling, Understanding, and Prediction," *Proceedings of the IEEE Workshop on Motion: Representation and Analysis,* Kiawah Island, S.C., 1986, pp. 125–130.

T. S. Huang, S. D. Blostein, and E. A. Margerum, "Least-Squares Estimation of Motion Parameters from 3-D Point Correspondences," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1986.

IEEE Comput. Mag., Special issue on Computer Analysis of Time-Varying Images **14**(8) pp. 7–69 (Aug. 1981).

IEEE Trans. PAMI, Special issue on Motion and Time-Varying Imagery **2**(6), 493–588 (Nov. 1980).

K. Kanatani, "Tracing Planar Surface Motion from a Projection Without Knowing the Correspondence," *Comput. Vis. Graph. Img. Proc.* **29,** 1–12 (1985).

K. Kanatani, "Detecting the Motion of a Planar Surface by Line and Surface Integrals," *Comput. Vis. Graph. Img. Proc.* **29,** 13–22 (1985).

J. J. Koenderink and A. J. Van Doorn, *Depth and Shape from Differential Perspective in the Presence of Bending Deformation,* Preprint, Department of Medical and Physiological Physics, Princetonpiein 5, Utrecht, The Netherlands, 1985.

R. Kories and G. Zimmermann, "A Versatile Method for the Estimation of Displacement Vector Fields from Image Sequences," *Proceedings of the IEEE Workshop on Motion,* 1986.

D. T. Lawton, "Processing Translational Motion Sequences," *Comput. Vis. Graph. Img. Proc.* **22,** 116–144 (1983).

M. D. Levine, P. B. Nobel, and Y. M. Youssef, "A Rule-Based System for Characterizing Blood Cell Motion," in Huang, 1983.

J. Limb and J. Murphy, "Estimating the Velocity of Moving Images in TV Signals," *Comput. Graph. Img. Proc.* **4,** 311–327 (1975).

Y. C. Liu and T. S. Huang, "Estimation of Rigid Body Motion Using Straight-line Correspondences," *Proceedings of the IEEE Workshop on Motion: Representation and Analysis,* 1986, pp. 47–52.

H. C. Longuet-Higgins, "A Computer Program for Reconstructing a Scene from Two Projections," *Nature* **293,** 133–135 (Sept. 1981).

H. C. Longuet-Higgins, "The Visual Ambiguity of a Moving Plane," *Proc. Roy. Soc. Series B* **223**(1), 165–170 (1984).

J. C. Longuet-Higgins, "The Reconstruction of a Scene from Two Projections-Configurations that Defeat the 8-Point Algorithm," *Proceedings of the First Conference on Artificial Intelligence Applications,* Denver, Colo., 1984, pp. 395–397.

A. Mitiche and J. K. Aggarwal, "A Computational Analysis of Time-Varying Images," in T. Y. Young and K. S. Fu, eds., *Handbook of Pattern Recognition and Image Processing,* Academic Press, New York, 1985.

A. Mitiche, S. Seida, and J. K. Aggarwal, "Line-Based Computation of Structure and Motion Using Angular Invariance," *Proceedings of the IEEE Workshop on Motion: Representation and Analysis,* 1986, pp. 175–180.

H. P. Moravec, "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Ph.D dissertation, Stanford University, September 1980.

H. H. Nagel, "Displacement Vectors Derived from 2nd-Order Intensity Variations in Image Sequences," *Comput. Vis. Graph. Img. Proc.* **21,** 85–117 (January 1983).

H. H. Nagel, "Constraints for the Estimation of Displacement Vector Fields from Image Sequences," *Proc. of the Eighth IJCAI,* Karlsruhe, FRG, Morgan-Kaufmann, San Mateo, Calif., Aug. 8–12, 1983, pp. 945–951.

K. Nakayama, "Biological Image Motion Processing: A Review," *Vis. Res.* **25,** 625–660 (1985).

R. C. Nelson and J. Aloimonos, "Finding Motion Parameters from Spherical Flow Fields (Or the Advantages of Having Eyes in the Back of Your Head)," *Biological Cybernetics* **58,** 261–273 (1988).

B. Neumann, *Natural Language Description of Time-Varying Scenes,* Bericht No. 105, FBI-HH-B-105/84, Fachberich Informatik, University of Hamburg, FRG, 1984.

J. O'Rourke and N. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation," *IEEE Trans. PAMI* **2,** 522–536 (1980).

J. A. Orr, D. Cyganski, and R. Vaz, "Determination of Affine Transforms from Object Contours With No Point Correspondence Information," *Proceedings of the ICASSP 85,* 1985, pp. 24.10.1–4.

K. Prazdny, "Egomotion and Relative Depth Map from Optical Flow," *Biol. Cybernet.* **36,** 87–102 (1980).

Proceedings of the ACM Workshop on Motion: Representation and Perception, Toronto, 1983.

Proceedings of the IEEE Workshop on Motion: Representation and Analysis, Kiawah Island, S.C., May 7–9, 1986.

Proceedings of the Workshop on Computer Analysis of Time-Varying Imagery, Abstracts, University of Pennsylvania, Philadelphia, Pa., 1979.

J. D. Robbins and A. N. Netravali, "Recursive Motion Compensation: A Review," in Huang, 1983.

F. Rocca, "TV Bandwidth Compression Utilizing Frame-to-Frame Correlation and Movement Compensation," in T. S. Huang and O. J. Tretiak, eds., *Picture Bandwidth Compression,* Gordon and Breach, London, 1972.

M. E. Spetsakis, "The Geometry and Statistics of Visual Motion," Ph.D. dissertation, Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, 1989.

M. E. Spetsakis and J. Aloimonos, "Structure From Motion Using Line Correspondences," *Int'l. J. Computer Vision* **4,** 171–183 (1990).

M. E. Spetsakis and J. Aloimonos, "On Optimal Motion Algorithms," *Proceedings of the IEEE Workshop on Visual Motion,* Irvine, Calif., March 1989.

W. B. Thompson and S. T. Barnard, "Low-Level Estimation and Interpretation of Visual Motion," *IEEE Comput.* **14,** 47–56 (1981).

R. Y. Tsai and T. S. Huang, "Uniqueness and Estimation of 3-D Motion Parameters of Rigid Bodies with Curved Surfaces," *IEEE Trans. PAMI* **6**(1), 13–27 (Jan. 1984).

R. Y. Tsai and T. S. Huang, "Estimating 3-D Motion Parameters of a Rigid Planar Patch," *IEEE Trans. ASSP* **29**(7), 1147–1152 (December 1981).

R. Y. Tsai, T. S. Huang, and W. L. Zhu, "Estimating 3-D Motion Parameters of a Rigid Planar Patch. II: Singular Value Decomposition," *IEEE Trans. ASSP* **30**(4) (Aug. 1982); correction: **31**(2), 514 (April 1983).

R. Y. Tsai and T. S. Huang, "Estimating 3-D Motion Parameters of a Rigid Planar Patch. III: Finite Point Correspondences and the Three-View Problem," *IEEE Trans. ASSP* **32**(2), 213–220 (April 1984).

J. K. Tsotsos, J. Mylopoulos, H. D. Corvey, and S. W. Zucker, "A Framework for Visual Motion Understanding," *IEEE Trans. PAMI,* **2**(6), 563–573 (November 1980).

S. Ullman, *The Intepretation of Visual Motion,* MIT Press, Cambridge, Mass., 1979.

S. Ullman, "The Interpretation of Visual Motion," Ph.D. dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1979.

S. Ullman, "Analysis of Visual Motion by Biological and Computer Vision Systems," *IEEE Comput.* **14,** 57–69 (1981).

A. Verri and T. Poggio, "Against Quantitative Optic Flow," *Proc. IEEE International Conference on Computer Vision,* 1987.

A. M. Waxman and S. Ullman, *Surface Structure and 3-D Motion from Image Flow: A Kinematic Analysis,* CAR-TR-24, CS-TR-1332. Center for Automation Research, University of Maryland, College Park, October 1983.

A. M. Waxman and K. Wohn, *Contour Evolution, Neighborhood Deformation, and Global Image Flow: Planar Surface in Motion,* CAR-TR-58, CS-TR-1394, Center for Automation Research, University of Maryland, College Park, April 1984a.

A. M. Waxman and K. Wohn, "Contour Evolution, Neighborhood Deformation and Image Flow: Textured Surfaces in Motion," in W. Richards and S. Ullman, eds., *Image Understanding '84,* Ablex, Norwood, NJ, 1984b.

J. A. Webb and J. K. Aggarwal, "Structure from Motion of Rigid and Jointed Objects," *Artif. Intell.* **19**(1), 107–130 (1983a).

J. A. Webb and J. K. Aggarwal, "Shape and Correspondence," *Comput. Vis. Graph. Img. Proc.* **21,** 145–160 (1983b).

B. L. Yen and T. S. Huang, "Determining 3-D Motion and Structure of a Rigid Body Using Straight Line Correspondences," in Huang, 1983, pp. 365–394.

B. L. Yen and T. S. Huang, "Determining 3-D Motion Parameters of a Rigid Body: A Vector-Geometric Approach," *Proceedings of the ACM Workshop on Motion,* Toronto, 1983.

T. Y. Young and Y. L. Wang, "Analysis of 3-D Rotation and Linear Shape Changes," *Patt. Recog. Lett.* **2,** 239–242 (1984).

X. Zhuang, T. S. Huang, and R. M. Haralick, "Two-view Motion Analysis: A Unified Algorithm," *J. Opt. Soc. Am.* **3**(9), 1492–1500 (Sept. 1986).

X. Zhuang, R. M. Haralick, and J. S. Lee, "Rigid Body Motion and the Optic Flow Under a Small Perturbation," *IEEE Trans. PAMI* (in press).

T. S. HUANG
University of Illinois

The preparation of this article was supported by Scientific Services Program, Battelle Columbus Laboratories contract DAAG29-81-D-0100. This article is adapted and updated from the articles in the first edition entitled "Motion Analysis," by T. S. Huang, and "Optical Flow," by E. Hildreth of MIT. The help of J. (Y.) Aloimonos in carrying out the adaptation and updating of this article is gratefully acknowledged.

VISUAL PERCEPTION

The veridicality of visual perception, that is, the correspondence between the percept of the outside world and its physical features that can be verified, eg, through the sense of touch, is probably the most striking quality of vision. Despite its numerous shortcomings, catalogued and studied by psychologists as visual illusions, the visual system virtually never fails to provide information about the outside world that is of genuine behavioral importance. The three-dimensional layout of surfaces in the vicinity of the observer and the motion, compositions, and grouping of these surfaces into well-defined objects are representative examples. The resulting impression of the visual world is complete; the visual experience of the surrounding space has no gaps in it, even in those portions of it for which no input information is available (such as that part of the space projected onto the blind spot of the eye). At the same time, through fixation, attention, and scrutiny we can perceive many fine details of the environment.

QUESTIONS

Historically, reflections of the sort outlined above have led to many questions pertaining to the phenomenon of perception. These questions can be divided into distinct classes, according to the aspects of vision they address. This essay will deal with only a few of these aspects. It will survey some common features of the human perceptual performance, the processes involved in that performance and how these are studied experimentally. [Boff and co-workers (1986) is a recent comprehensive handbook on perceptual performance.] Computational accounts of vision (Marr, 1982), also offered elsewhere in the encyclopedia, will be pointed to briefly. Finally, most of the topics pertaining to the biological substrate of vision, except some of the most pervasive notions (such as receptive fields and cortical maps), will be omitted altogether. For information on the anatomy and the physiology of vision the reader is referred to other sources (e.g., Kandel and Schwartz, 1985).

Visual Performance

Gibson (1950) begins his book on visual perception with a question: "Why do things look as they do?" An obvious (but not very informative) answer to this would be: "because of the way our visual system is built." In fact, the recounting of conditions that must be fulfilled before anyone can see, which constitutes the first paragraph of Gibson's book, begs for a more constructive formulation of the basic empirical question of visual perception: how does the visual system perform under different conditions, or, what are the factors that affect the way things look? Some of the factors that influence perceptual performance, such as the basic architecture of the visual system, are internal to the observer and belong to the domain of visual neuroscience. Other factors, such as the physical characteristics of the stimulus and of the manner of its presentation, can be externally manipulated. Experimental psychological study of visual perception is aimed at understanding the outcome of the perceptual process, its building blocks, and its neural substrate, through controlled manipulation of the external stimulus.

Consider as an example Figure 1a. When asked to describe its contents, average observers usually state that it shows two overlapping triangles, of which one has its vertices at the centers of the black disks and consists predominantly of illusory contours. Why do we see this second triangle? One can make an initial step toward an answer by trying to influence the percept of the illusory triangle by manipulating the stimulus (Fig. 1b). As a by-product, this process of perceptual experimentation frequently comes up with notions that are useful for understanding other visual phenomena (in this specific example, the important notions are those of completion and filling in, to be discussed later), including the perception of natural scenes.

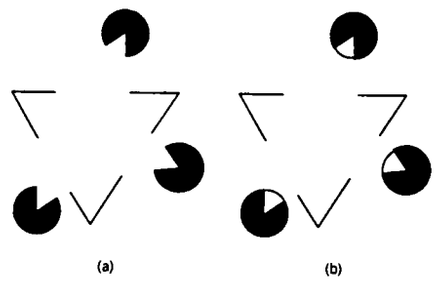


Figure 1. (*a*) Illusory contours, as well as some abnormal depth and lightness percepts, arise in this version of Kanizsa's triangle; (*b*) a modified version, in which the illusory percepts are much weaker.

Visual Competence

A different angle on the problem of perception is provided by the opening lines of Marr's (1982) book on computational vision. Marr asks "What does it mean to see?" This question, which (using Chomsky's terminology) is about competence rather than performance, cannot be elaborated on much further without making a commitment to an important part of the answer. One of the two main approaches to this problem is through the notion of direct perception (Gibson, 1979). According to Gibson, perception is direct because the organism picks up relevant perceptual invariants, such as the three-dimensional shape of objects, from the visual world without intervening processing or representation. (For a long time, Gibson went so far as to deny the existence of retinal images and the relevance to perception of geometric optics involved in their formation.) In its rejection of internal representations and processes that can be described as unconscious inference, the direct perception stance is related to behaviorism. The answer of Gibson's school to the question "What does it mean to see" is "to be attuned to certain invariant qualities present in the optic array, and, potentially, to become disposed to act in certain ways, given the appropriate stimulation." Under this view, phenomena such as that illustrated in Figure 1 tend to be dismissed as unnatural and "ecologically invalid."

The other approach, called "representational" by Marr (1982), postulates that the goal of any visual system, including biological vision, is to produce representations of the environment, and that various specific visual tasks, such as recognition and navigation, are solved through inference, or computation, defined over these representations (Marr, 1982). The notion of perception as unconscious inference is usually traced back to von Helmholtz (1856/1964). In his treatise on physiological optics von Helmholtz argued that most often the information present in the visual world as projected on the eye's retina is too incomplete to support the richness of perception, which perforce must largely rely on previously or independently available data and on unconscious reasoning. According to Marr, an important source of this additional information is a knowledge of physics of the outside world and of image formation.

Consider again the example of Figure 1. A computational account of the perceived illusory triangle would indicate that (1) all contours, including the illusory ones, are explicitly represented at some stage of the visual system, and (2) the transformation between the stimulus and

the explicit representation can be computationally specified (and implemented in the hardware of the brain). A distinct advantage of computational theories of perception is that, unlike "direct" accounts, they can be made to generate concrete predictions, which, in turn, can be experimentally upheld (or refuted). Computational, psychological, and physiological investigations of illusory contours are reported in Ullman (1976), Cavanagh (1987), and von der Heydt and co-workers (1984), respectively.

To summarize, the interesting questions that can be asked about visual perception are those of performance (what conditions affect the way we see?) and of competence (what does it mean to see? how do we see?). The computational–representational paradigm proved useful in addressing all these questions. At present this approach dominates the psychology of vision, consequently; in the remainder of the article its assumptions and terminology are employed without qualification.

THE STUDY OF PERCEPTION

Goals

One way to begin investigating the basic questions mentioned above is to decide on the form into which the answers should be cast. As in other natural sciences, this calls for the formation of a mathematical framework that would encompass the existing body of data on perceptual performance and would admit generalization by successfully predicting performance under novel conditions. In relatively primitive organisms the process of perception appears indeed to be amenable to precise mathematical treatment. As Poggio and Reichardt (1976) have shown, in the fly the transfer function of the module that supplies visual motion information for the purpose of flight control can be specified by a simple expression with only a few terms. The situation in human perception is rather more complicated. Among other reasons, this is because of the sheer diversity and complexity of the human visual system, because the output representations of visual modules are as yet unknown, and because in many cases top-down influences tend to interfere with "pure" perception. A typical example of the classical mathematics of perception is Luce (1986), where one can find an exhaustive survey of the response time paradigm (see the section on methods below). The main mathematical tools of this paradigm are descriptive statistics and statistical models borrowed from signal detection theory (Green and Swets, 1966).

Shepard's proposal of a "universal law of generalization for psychological science" (Shepard, 1987) represents a more ambitious attempt at the mathematization of perception. The article, written for the occasion of the tercentenary of the publication of Newton's *Principia Mathematica*, suggests that the generalization of response from a learned to a novel stimulus depends on the distance between the stimuli in an abstract space that has the same metric structure for a wide variety of tasks, ranging from shape, size, and color judgments to auditory signal perception. While the notion of a general psychological space may apply to stimuli that themselves possess a well-defined metric structure, the chances are meager that the more cognitive perceptual processes such as object recognition would admit a similar universal law-like or nomological description. If those philosophers are right who claim that mental processes are anomalous instead of nomological (Davidson, 1980), accounts of cognition in terms of prototypes and narrow-scope or local rules should be more fitting than invocations of universal laws.

Methods

There are many ways to collect the data necessary for building a theory of perception. One may distinguish between psychological approaches, which concentrate on the perceptual capacities and experiential aspects of perception, and biological approaches, which focus on the anatomy and physiology of the sensory nervous system. Only one approach is discussed here: experimental psychology.

Experiments that quantify and measure the psychometric function (viz, the response of a subject to a controlled stimulus) have traditionally been the principal method of the psychological study of perception. This experimental paradigm, called psychophysical because it relates the magnitude of a psychological response variable such as response time to some physical quality of the stimulus, dates back to the previous century. A clear formulation, due to Jastrow (1890), states that if the process of perception is indeed structured, then different paths through this structure will yield different response times. If this is true, one may hope to infer the structure of perception from the patterns of response times obtained under different experimental conditions. Mental rotation [see Shepard and Cooper (1982) for an overview and Pylyshyn (1985) for a critique] provides an outstanding example of a phenomenon in which the dependence of response time on a characteristic of the stimulus has triggered hundreds of experiments and was incorporated into the foundations of a theory of visual representation. (In this case, the task called for a judgment of object identity between two simultaneously shown images of three-dimensional objects, and the response time was found to depend linearly on the relative misorientation of the objects with respect to one another.)

BUILDING BLOCKS

Reductionistic methods that investigate the structure of the perceptual system encourage the dissection of vision into submodalities. Some of these building blocks of perception, such as lightness, hue, texture, stereopsis, visual motion, the perception of space, and object recognition, are briefly described below. Within the scope of this article, little more than a hint can be given as to the perceptual problems solved by each module.

Lightness and Shading

Confronted with a gradient of illumination across the viewed surface, the visual system must separate the effect of illumination from the effects of surface albedo, orientation, and shape. Disentangling all the factors that contribute to the intensity of the image at a given point on the

retina is probably the most complicated computational problem in vision. Thus it is not surprising that some of the more compelling illusions, such as the Mach bands and the Craik–Cornsweet–O'Brien effect [see, eg, Frisby (1979) for an overview], are caused by peculiarities of lightness perception mechanisms (one such peculiarity is illustrated in Fig. 1a, where the illusory triangle is perceived to be brighter than the background). In addressing the lightness problem, the human visual system appears to have settled for qualitative rather than quantitative solutions (Todd and Mingolla, 1983). Moreover, these solutions often seem to be based on high-level heuristics (Ramachandran, 1988) and are easily downplayed if the relevant information is available from other sources, such as the shape of the occluding contours of the surface (Koenderink, 1984) or stereopsis (Bülthoff and Mallot, 1988).

Color

In the perception of color, as in the perception of lightness, the human visual system exhibits an impressive disregard for irrelevant variables. In this case, the intensity and the spectral content of the illuminant must be factored out if a reasonable approximation to the color of the viewed surface is to be inferred [see the review in Boynton (1978)]. The mechanisms responsible for this function appear to be similar for lightness and for color in that they depend on local contrast while ignoring slow and gradual changes in image intensity and spectrum, which in many cases can be safely attributed to the influence of the illuminant (Land and McCann, 1971). Observations that are not easily accounted for by an application of such simple fixed rules were made by Gilchrist (1977), who found that global (and cognitive) factors such as the knowledge of the spatial arrangement of surfaces may affect their perceived lightness. (See also COLOR VISION.)

Texture

In the natural world, texture, along with color, is an important cue to the physical composition of visible surfaces and can be used to segment complex scenes into surface patches that have distinct origins (eg, belong to different objects). Also, texture gradients can be readily interpreted in terms of the orientation of the underlying surfaces and thus contain cues to the three-dimensional structure of the visual space (Gibson, 1950). Texture is a mass phenomenon; a surface must bear more than a few markings to be perceived as textured. The problem of texture perception can be formulated either in statistical terms, or in terms of the detection of the underlying texture elements or textons (see Julesz, 1984). An issue that was originally raised in the context of texture perception and has since been intensively studied is that of preattentive discrimination (ie, distinguishing between different stimuli without the involvement of attention or scrutiny). Presumably, features that combine into preattentively discriminable textures are processed in parallel over the entire visual field. The identification of such features provides important clues to the structure of early visual processes. An illustration of this approach may be found in recent work by Fahle (1990), who found that vernier offset stimuli in the hyperacuity range (pairs of abutting line segments displaced by an amount that is smaller than the spacing of the photoreceptors in the retina) can be detected in parallel. (See also TEXTURE.)

Stereopsis

Since the retinal image is a projection of the three-dimensional world onto a two-dimensional surface, the information on the third dimension, depth, is already lost at the very first stage of vision. The perception of depth, or stereopsis, can, however, still be attained by combining information from the two eyes. (Monocular depth cues are mentioned below in the section on space perception.) Stereopsis works because the separation between the retinal images of objects (retinal disparity) is different in the left and right eyes, depending on the separation of the objects in depth. Binocular stereo resolution is extremely fine: a difference in depth of 1 mm can be perceived at a distance of 1 meter (m). The disparity between the two eyes' views under such conditions is many times smaller than the size of a single retinal photoreceptor.

Stereopsis has received much attention in the study of vision [Julesz (1971); Poggio and Poggio (1984); see also STEREO VISION]. The behavioral importance of depth perception becomes apparent if one attempts to thread a needle, or catch a fly, with one eye closed. Stereopsis is an acquired ability: newborn babies do not perceive binocular depth until the age of 3 or 4 months. Disorders such as strabismus (the inability to fixate the same object simultaneously with both eyes) present during this period of plasticity can cause permanent stereoblindness by hampering the development of one of the two main processes underlying stereopsis: matching the two retinal images to produce the disparity field. The power of the matching process is illustrated by our ability to perceive depth in random-dot stereograms [image pairs consisting of random dots, some of which are displaced in one image with respect to the other to form a stimulus that can be perceived only through stereo vision; see Julesz (1971)]. In a random-dot stereogram, each dot in one image can match potentially any dot in the other image. Although matching is usually sufficiently selective to disambiguate such situations, in some cases people may perceive simultaneously several surfaces corresponding to multiple matchings between elements in the two images (Weinshall, 1989). Moreover, the relationship between the outcome of the matching and the perceived depth is sometimes not unequivocal. For example, the perceived depth may correspond to an average disparity rather than to one of the actual disparities derived from a possible matching (Mitchison and McKee, 1985).

The second process involved in stereopsis is surface interpolation, which fills in the gaps between those locations in the image where exact disparities are found through matching. Similar to matching, the surface interpolation subsystem possesses several distinct features that are not well understood and are not reproduced by machine vision algorithms. Two of these are simultaneous perception of

multiple transparent surfaces and the integration of different depth cues in surface perception.

Motion

Visual motion contains many important cues about the outside world (see also VISUAL MOTION ANALYSIS). Moving patterns of light projected onto the retina provide information that can be used to segment the surrounding scene into objects according to their motion and to navigate in the environment while avoiding collisions with both stationary and moving obstacles. Visual motion can also be interpreted to yield the three-dimensional structure of objects (Wallach and O'Connell, 1953), even when the objects themselves are allowed to deform while moving, a common phenomenon in the motion of living things (see Johansson, 1973). The autonomy of visual motion perception is demonstrated by our ability to perceive three-dimensional structure in moving two-dimensional patterns, such as those that appear on a television screen, and even in random-dot kinematograms (Ullman, 1979). The contribution of motion to our overall impression of the world can be appreciated by anyone who has watched a film taken from the vantage point of a roller-coaster rider: the somatic illusions evoked by such stimuli are strong enough to override vestibular and somatosensory cues.

As Rock (1984) has pointed out, the presence of retinal displacement of objects is neither sufficient nor necessary for the perception of motion, despite the indisputable fact that such displacement is the starting point of neural motion processing. On one hand, moving objects that are fixated and tracked are effectively stationary with respect to the retina, but are still perceived as moving. On the other hand, we perceive the visual world as immobile when our eyes move between fixations (but not when the eyes are moved externally, eg, by gently pressing on the eyeball from the side).

Three-Dimensional Space and Shape

The perception of the three-dimensional layout of visual space and of solid objects embedded in it relies on a combined action of all the basic modules mentioned above, as well as on a variety of perceptual rules of thumb that cannot be readily attributed to one of those modules (Ramachandran, 1988). The role of shading, textural cues, retinal disparity, and motion information in seeing depth has been outlined above. Another class of depth cues is provided by the oculomotor system; depth can be estimated from convergence of the eyes and from the accommodation status of the lens. Among the pictorial cues are interposition (inferred from occlusion of some objects by others), shadows, perspective, and familiar size information (see Rock, 1984). The problem of understanding the integration process that brings the depth cues together, traditionally neglected in favor of the study of individual visual modules, has recently received increasing attention (Bülthoff and Mallot, 1988). The main aspects of the integration problem are the nature of the output representation and the relative weight given to each cue. Situations in which the cues are conflicting can be especially interesting. For example, when the contents of Figure 1a are shown stereoscopically in such a manner that the vertices of the illusory triangle appear, in conflict with the (imaginary) interposition cue, behind those of the real one, the illusion becomes weaker (Gregory, 1978).

Object Recognition

Mechanisms that support object recognition (qv) in human vision are the subject of considerable controversy among psychologists. Although most would agree that recognition involves comparison between the stimulus and an internal model or representation kept in memory, no consensus exists as to the nature of that representation. Consider the problems encountered by the visual system that attempts to identify an object present in the field of view. Assuming that the candidate object has already been detected and its approximate location estimated, the system must segment the object from the environment and factor out variations in its appearance due to changing illumination, changing viewpoint (see Fig. 2), and, possibly, changing shape of the object (as in the recognition of a moving animal; see the section on motion above).

Does the visual system represent objects in a straightforward fashion, eg, by remembering sets of two-dimensional "snapshots" taken from different vantage points, or are the object models, geometrically, three-dimensional analogs of the objects they represent? Arguments based, among other phenomena, on our ability to perceive and describe the three-dimensional shape of novel objects led many researchers [notably Marr (1982)] to postulate the formation of three-dimensional object-centered (ie, viewpoint-invariant) representations of the environment to be the ultimate goal and the final product of vision. This view amounts to much more than a theory of object recognition; it dictates the interpretation of processes of early vision in terms of the reconstruction of a replica of the visual world. Recently, this view has been disputed on philosophical, empirical and computational grounds (see Sloman, 1987; Quinlan, in press; Edelman and Weinshall, 1989; Edelman and Bülthoff, in press).

Three-dimensional object-centered models, envisaged by Marr, and other three-dimensional structural representations (eg, Biederman, 1985) are only a few of the theories competing in the field of recognition. Major alternatives (see Ullman, 1989, for a review) are template matching, description by invariant features and shape

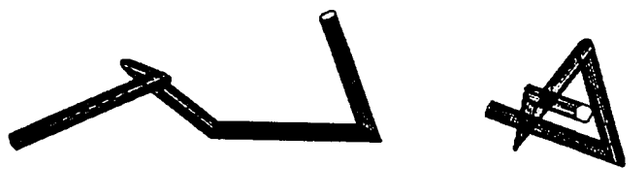


Figure 2. The appearance of a three-dimensional object can depend strongly on the viewpoint. The image on the right is of the same object as the image on the left, rotated in depth by 90°. People find it difficult to recognize such objects from a radically unfamiliar viewpoint, even when stereo information (Rock and DiVita, 1987) or both stereo and motion cues (Edelman and Bülthoff, 1990) to the three-dimensional shape of the object are available.

normalization. Although it is clear that structural descriptions of the type suggested by Marr, Biederman, and others must be invoked to explain some perceptual phenomena, the emergence of recognition models based on interpolation among prototypical two-dimensional views (e.g., Ullman and Basri, 1990; Poggio and Edelman, 1990) indicates that memory for specific instances may be more important for recognition than previously believed.

CROSS-MODAL CHARACTERISTICS AND PROCESSES

This section lists several characteristics of perception whose common denominator is generality and pervasiveness in vision. These are grouped into two classes. The first class includes phenomena that are common to more than one of the building blocks mentioned earlier. The other class comprises dynamic processes whose scope spans several visual submodalities.

Constancy

Perceptual constancy is our tendency to see properties of objects as invariant despite perpetually changing retinal stimuli. Following is a list of the most prominent constancy phenomena. (Some of these have already been mentioned in the preceding sections.) Concrete examples of each phenomenon can be found, eg, in Rock (1984).

Lightness Constancy. The perception of the shade of a surface's lightness varies in general with the true albedo of the surface rather than with its luminance (the intensity of the light reflected by the surface, which changes, for example, because of varying illumination and shadows).

Color Constancy. When the spectral content of the illuminant undergoes a radical change, a surface will no longer reflect light that corresponds to its true color (ie, the color it reflects when illuminated with white light). Nevertheless, its perceived color will in general appear close to the true one.

Size Constancy. A given object in the world appears to be about the same size, irrespective of the variation of the size of its retinal projection (due, eg, to its varying distance from the eye).

Shape Constancy. The perceived shape of an object remains constant despite changes in the shape of its retinal projection caused by the movement of the object relative to the observer.

Space Constancy. The visual world appears to us as stable and unmoving despite continuing movement of the retinal image, caused by the movement of the eyes, as in visual tracking and saccades (see the section on attention and search below), the head and the entire body (as in locomotion).

Explanations of Constancy. The constancies of visual perception constitute an essential part of the visual experience as we know it. Imagine what a bewildering world it would be if red tomatoes turned yellow when viewed in the kitchen under incandescent light, coins appeared elliptical unless viewed from the proper vantage point, and the slightest movement of the head sent the entire surrounding scene careening about. How does the visual system achieve perceptual constancy? The stimulus-relation explanation, favored by the direct perception school, sees the stimulus itself as the sole cause of constancy. According to this account, the context in which the stimulus appears affects the way it is perceived. For example, the apparent size of an unfamiliar object may be affected by the presentation next to it of another object of known size. In many situations, however, context information is unavailable, yet size constancy still holds. To continue the preceding example, the size of a luminous object in a darkened room may be correctly judged as long as distance information is available, eg, from accommodation and convergence. If, however, the object is seen through a narrow aperture that eliminates distance cues, the constancy breaks down and the object appears to be of indeterminate size (Rock, 1984). This phenomenon prompts an alternative explanation of constancy that proposes that viewers take unconsciously into account independent or prior knowledge relevant to a given situation. As we shall see in the next section, the involvement of prior knowledge can account also for other features of vision besides constancy.

Implicit Knowledge

An example of the facilitation of perception through the use of information not available in the immediate sensory input is the visual system's superior performance in recognizing and remembering objects and scenes that make sense, as opposed to those that do not (Biederman and coworkers, 1974; Potter, 1975). Although the implicit knowledge in this case is used unconsciously, the perceiver would normally be able to identify and describe its source. In other cases the knowledge source is not so readily apparent to the observer. For example, the visual system seems to take into account the physics of specular reflection in the perception of three-dimensional shape in shaded images (Blake and Bülthoff, 1990). In another example, motion perception appears to involve implicit familiarity with the physics of transparency (Stoner and coworkers, 1990).

Illusions

The unconscious rules of thumb that are in part responsible for the incredibly rapid performance of the visual system in a wide variety of vital tasks are, necessarily, limited in scope. Certain visual stimuli produce illusory or non-veridical percepts by causing the breakdown of those perceptual processes that rely on unconscious rules. Another source of illusions is in the inherent anatomical and physiological properties and limitations of the visual system. The illusory contours of Figure 1**a** for which a tentative physiological mechanism has been identified (von der Heydt and co-workers, 1984) are an example of this kind (see also Gregory, 1978).

Processes

It is remarkable that a number of central parts of the process of visual perception can be described functionally in a manner that is largely independent of the particular goals they serve. What follows here is an attempt to briefly characterize these subprocesses.

Adaptation. The most common example of adaptation is the adjustment in the light sensitivity of the eye that follows any change in the ambient illumination. The dynamic range of light adaptation is very wide (at least five orders of magnitude). Only a small part of this range (a factor of 16 or so) is attributable to the changes in pupil size; the rest is supported by physiological processes that operate at the photoreceptors and in the higher levels of the retina. Presumably, some of the visual aftereffects that may be classified as adaptation phenomena happen at higher levels of the visual system. A well-known example of this type is the motion aftereffect; we tend to perceive the stationary parts of a waterfall scene as moving upward after having concentrated for sufficient time on the downward motion of the falling water. One possible explanation of the aftereffects is that the visual system represents qualities such as planarity and immobility as a dynamic balance between representations of opposites such as convexity–concavity and upward–downward motion. A fatiguelike reduction in the activity level of the representation substrate of downward motion, for example, will then cause a perception of upward motion in a stimulus that is, in fact, static. Note that this account relies on another common property of perception related to adaptation: the preferential response to temporal change and spatial contrast, as opposed to status quo and uniformity.

Attention and Search. The visual system allocates different amounts of processing resources to different portions of the visual field. Most of this nonuniformity is architectural and is due to the gradual decrease in the photoreceptor density (which entails a decrease in acuity) between the fovea and the periphery. The other part is more flexible and can be manipulated at will. The phenomenon of diverting processing resources to specified locations in the visual field or to a specified submodality is called visual attention (see, eg, Keele and Neill, 1978). Attention can be shifted to a new location overtly by executing an appropriate saccadic eye movement or covertly by a mechanism whose precise nature is not yet known. The benefit of the ability to shift attention is in the economy of processing resources required for adequate functioning. For some visual operations, such as search for compound stimuli (Treisman and Gelade, 1980), maintaining uniform processing capability over the entire field may be quite expensive. In these cases, attention provides an acceptable trade-off between resources and time.

Perceptual Organization. Perceptual organization is a collective term for a diverse set of processes that contribute to the emergence of order in the visual input. Some of the phenomena already mentioned here can be considered as particular instances of perceptual organization. Two examples that are, in reality, opposite sides of the same coin are shape constancy and visual motion. The perception of a deforming two-dimensional retinal stimulus pattern as a three-dimensional object in motion amounts to organizing the visual input so that it can be described in a simple and stable fashion. The study of perceptual organization has a long history in psychology. Palmer (1983), following Cassirer, Pitts and McCulloch, Gibson, and others, outlined a uniform framework for the study of organizational phenomena, based on the mathematical notion of invariance under transformation groups. The transformational approach allows problems of constancy and motion to be addressed in the same language as the classical issues of perceptual organization: figural goodness, grouping and frames of reference. A major attempt to understand these issues, motivated by a conviction that such understanding would shed light on perception in general, led to the formation of the Gestalt school in psychology (Köhler, 1947). Many of the laws of organization proposed by Gestalt theorists, as well as the concept of perceptual goodness (*Praegnanz*) they have introduced to account for a variety of perceptual phenomena, have been incorporated in a more rigorous formulation into the currently prevailing paradigm of visual perception [see the discussion in Marr (1982), p. 187].

Completion and Filling in. The group of phenomena that can be characterized by a tendency to optimize figural goodness includes two that have been mentioned in the section on visual competence (see also Figure 1). The first of these is contour completion: the visual system prefers to perceive a nonexistent contour forming the illusory triangle rather than see the three dented disks as unrelated to each other (in which case the missing portions of the real triangle would remain unaccounted for). It turns out that the perceived shape of the completed contour can be produced by a process that minimizes a measure of its curvature (Ullman, 1976). The second phenomenon that affects the perception of figures as wholes is that of area filling in, which may be considered a two-dimensional analog of contour completion. In Figure 1, filling in is apparent in the increased subjective lightness of the illusory triangle. Another manifestation of filling in is motion capture: our tendency to perceive stationary features that happen to fall within a moving contour as drifting along with the contour. Surface interpolation (see the section on stereopsis) is an instance of filling-in that, analogously to contour completion, can be formulated as a process of optimization (see Poggio and co-workers, 1985).

Categorization. Several illustrations of our disposition to see the environment as structured instead of chaotic were given in the previous sections, when we discussed the phenomena of object constancy and motion perception. The imposition of this kind of high-level structure on the visual world is an apotheosis of the processes of perceptual organization, linking vision to general cognition and language. Experimental evidence suggests that this connection is bilateral, and that the cognitive level can influence visual perception. The most directly relevant experiments

are those in which subjects exhibit object superiority effects; an example is the facilitation of the perception of a low-level feature, such as a line segment, by virtue of its appearance as a part of the projection of a three-dimensional object (Weisstein and Harris, 1974).

In many top-down effects (including object superiority), the perceptual phenomenon is better characterized as categorization than recognition. The manner of cognitive involvement in perception is thus more flexible and more general than mere recollection of previously encountered stimulus exemplars. Experiments carried out by Rosch and her collaborators as part of a wider study of the structure of categories (see, eg, Rosch and co-workers, 1976) showed that people tend to perceive and describe objects at a certain level of detail. Importantly, this basic category level can be independently defined in terms of visual perception, language, and general cognitive development.

Some of the features of the processes involved in perceptual learning and memory are mentioned in the next section.

Perceptual Learning. Perceptual learning, or the adjustment of perception to the stimulus aspects of the environment, is sometimes distinguished from cognitive learning; the latter term is reserved for the modification of problem-solving behavior (Walk, 1978). In early vision, learning occurs in processes such as adaptation, mentioned above. The famous experiment first made by Stratton (1897/1964), in which a subject wearing inverting prisms gradually adapts to this condition, is a forceful reminder that the degree of plasticity at lower levels of the visual system should not be underestimated. (See Rock, 1984, for a discussion of the inversion experiments.)

Exactly what is learned and what is innate in vision (and in cognition in general) has been the subject of intense philosophical debate since Plato's time [see Dretske (1990) for an overview]. A century of research in visual psychophysics and neurobiology of vision shows that the basic perceptual abilities of the human visual system (such as the ability to perceive luminance contrast) are largely innate, while others (such as some varieties of object constancy) are acquired and depend on the visual experience (Spelke, 1990). Significantly, the mechanisms of perceptual organization used by infants in learning how to see seem to persist through adulthood. Thus, the study of the ontogeny of visual perception may help clarify the nature of the long-term memory representations of objects and scenes.

A similar angle on the problem of representation is provided by studies in which the subjects' perceptual performance in three-dimensional object recognition is modified merely as a result of practice or exposure to the stimuli, without any feedback from the experimenter (after all, infants acquire vision, and language, without being instructed). Normally, the subjects' response time in recognition depends monotonically on the misorientation of the stimulus with respect to some canonical attitude [Palmer and co-workers (1981); see also the discussion of mental rotation in the section on methods, above). Perceptual learning under such circumstances, inferred from the observed changes in the pattern of performance (specifically, from increasingly uniform response times for different aspects of the stimulus), can be attributed to a shift towards a more memory-intensive and less time-consuming recognition strategy (Tarr and Pinker, 1989; Edelman and Bülthoff, 1990). Indeed, such a strategy appears to be the most suitable one for a system in which memory is cheap, but time is expensive.

THE NEURAL SUBSTRATE

The architecture of the human visual system reflects the major functional constraint imposed on it, namely, the requirement of being able to recognize or classify in a few hundreds of milliseconds any object from a potentially unlimited repertoire, while taking into account a variety of visual clues (Biederman and co-workers, 1974; cf Rosenfeld, 1987). The following section contains several hints as to what the basic functional elements that constitute this architecture may be.

Vision Is Massively Parallel. The parallelism of visual information processing becomes apparent already at the level of the retina, where a separation occurs between several pathways, each of which is functionally specialized to support certain aspects of the input, such as form, motion, and color. Since eventually these have to be integrated into a coherent percept, the cortical areas fed by the different pathways are interconnected in an orderly fashion [see recent reviews in Zeki and Shipp (1988) and Kaas (1989)], so that on the whole the architecture is a heterarchy in which lateral connections and shortcuts abound, rather than a hierarchy envisaged by early visual scientists. Remarkably, in all the visual areas, as in the entire neocortex in general, information is processed by the same variety of cells, arranged in the same columnar structure (Gilbert, 1988). Thus, any comprehensive theory of brain function would have to specify, in information-processing terms, what the basic operation is that can be supported by the cortical architecture and is at the same time sufficiently powerful to address the entire range of perceptual, cognitive, and motor tasks.

Maps and Receptive Fields. Although a viable and widely accepted theory of such scope has yet to be proposed, two well-established findings in the neurobiology of perception, and in particular of vision, provide an inkling as to the basic mode of information processing in the cortex. The first of these is concerned with the notion of the receptive field of a neuron. In vision, it is defined as that region of the visual field whose stimulation affects the activity of the neuron (see, eg, Kuffler and Nicholls, 1976). Without going into the details of the taxonomy and structure of receptive fields found in the visual cortex, one may describe their function schematically as the integration of information over a finite area of the visual field and its concentration at a single point: the axon of the neuron in question, or its output terminal. Note that the axon may, in turn, connect to many (potentially, tens of thousands) other neurons at a higher level, so that as a whole this structure is as heterachical as the large-scale arrangement of cortical areas mentioned above (Fig. 3a).

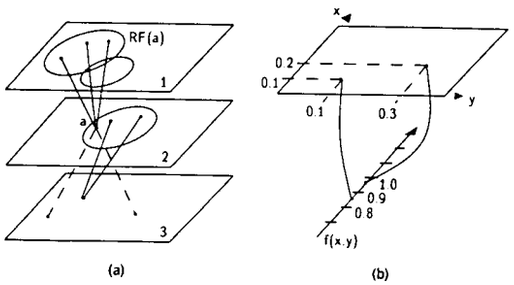


Figure 3. (*a*) A highly schematic illustration of the notion of a receptive field: unit a in area 2 receives input from the region marked as RF(a) in area 1, and, in turn, projects to many units in area 3; (*b*) computing the function $f(x) = e^{-(x^2+y^2)} \cos y$ with a two-dimensional–one dimensional (2D-to-1D) map of connections, or lookup table. Note that to find the value of the function at a point for which there in no entry in the lookup table one must resort to interpolation (see Poggio and Girosi, 1990).

The second basic notion in the architecture of vision, cortical mapping, pertains to the relationship among different receptive fields in the same visual cortical area. It turns out that many areas are interconnected by locally smooth maps (in fact, many areas are retinotopic, that is, their topology, but not necessarily their metric structure, conforms to that of the retina). Recent theoretical developments indicate that computing with maps or, equivalently, connections (Fig. 3b) is a powerful information-processing paradigm (see, eg, Ballard, 1986, for an integrative discussion). It has been suggested (Rojer and Schwartz, 1989; Mallot and co-workers, 1990; Damasio, 1989) that cortical mapping is the basic mechanism of the visual function of the brain. It remains to be seen whether this concept can be extended to encompass perception (and intelligence) in general.

CONCLUSION

The study of visual perception is bound to inspire awe, because of the recognition of the formidable problems posed by vision, and marvel, because of the appreciation of solutions developed by the brain to address these problems. Borrowing a phrase from Warren McCulloch, one can describe this study as discovering what a thing is, that a man may see it, and a man, that he may see things (McCulloch, 1965 p. 2). While looking for an answer to this question, it is worthwhile to remember that in perception, if not in intelligence, man is the measure of all things.

BIBLIOGRAPHY

D. H. Ballard, "Cortical Connections and Parallel Processing: Structure and Function," *Behav. Brain Sci.* **9,** 67–120 (1986).

I. Biederman, "Human Image Understanding: Recent Research and a Theory," *Comput. Vision, Graphics, Image Proc.* **32,** 29–73 (1985).

I. Biederman, J. C. Rabinowitz, A. L. Glass, and E. W. Stacy, "On the Information Extracted from a Glance at a Scene," *J. Exp. Psychol.* **103,** 597–600 (1974).

A. Blake and H. H. Bülthoff, "Does the Brain Know the Physics of Specular Reflection?," *Nature* **343,** 165–168 (1990).

K. R. Boff, L. Kaufman, and J. P. Thomas, eds., *Handbook of Perception and Human Performance*, Wiley, New York, 1986.

R. M. Boynton, "Color in Contour and Object Perception," in E. C. Carterette and M. P. Friedman, eds., *Handbook of Perception*, Vol. VIII, Academic Press, New York, pp. 173–199, 1978.

H. H. Bülthoff and H. A. Mallot, "Interaction of Depth Modules: Stereo and Shading," *J. Opt. Soc. Am.* **5,** 1749–1758 (1988).

P. Cavanagh, "Reconstructing the Third Dimension: Interactions between Color, Texture, Motion, Binocular Disparity and Shape," *Comput. Vision, Graphics, Image Proc.* **37,** 171–195 (1987).

A. R. Damasio, "The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones," *Neural Comput.* **1,** 123–132 (1989).

D. Davidson, *Essays on Actions and Events,* Clarendon Press, Oxford, 1980.

F. Dretske, "Seeing, Believing, and Knowing," in D. N. Osherson, S. M. Kosslyn, and J. M. Hollerbach, eds., *Visual Cognition and Action*, Vol. 2, MIT Press, Cambridge, Mass., pp. 129–148, 1990.

S. Edelman and H. H. Bülthoff, *Viewpoint-Specific Representations in 3-D Object Recognition,* AI Memo No. 1239, Artif. Intelligence Lab., MIT, Cambridge, Mass., August 1990

S. Edelman and D. Weinshall, "A Self-organizing Multiple-View Representation of 3D Objects," A.I. Memo No. 1146, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., August 1989.

S. Edelman and D. Weinshall, *Biol. Cybern.* **64,** 209–219 (1991).

J. P. Frisby, *Seeing*, Oxford University Press, Oxford, 1979.

J. J. Gibson, *The Perception of the Visual World*, Houghton Mifflin, Boston, 1950.

J. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, 1979.

C. D. Gilbert, "Neuronal and Synaptic Organization in the Cortex," in P. Rakic and W. Singer, eds., *Neurobiology of Neucortex*, Wiley, New York, pp. 219–240, 1988.

A. L. Gilchrist, "Perceived Lightness Depends on Perceived Spatial Arrangement," *Science* **195,** 185–187 (1977).

D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics,* Wiley, New York, 1966.

R. L. Gregory, "Illusions and Hallucinations," in E. C. Carterette and M. P. Friedman, eds., *Handbook of Perception*, Vol. IX, Academic Press, New York, pp. 337–357, 1978.

J. Jastrow, *The Time Relations of Mental Phenomena,* Hodges, New York, 1890.

G. Johansson, "Visual Perception of Biological Motion and a Model for Its Analysis," *Percept. Psychophys.* **14,** 201–211 (1973).

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

B. Julesz, "A Brief Outline of the Texton Theory of Human Vision," *Trends Neurosci.* **7,** 41–45 (1984).

J. H. Kaas, "Why Does the Brain Have so Many Visual Areas?," *J. Cogn. Neurosci.* **1,** 121–135 (1989).

E. R. Kandel and J. H. Schwartz, *Principles of Neural Science*, Elsevier, New York, 1985.

S. W. Keele and W. T. Neill, "Mechanisms of Attention," in E. C. Carterette and M. P. Friedman, eds., *Handbook of Perception*, Vol. IX, Academic Press, New York, pp. 3–47, 1978.

J. J. Koenderink, "What Does the Occluding Contour Tell Us about Solid Shape?," *Perception* **13,** 321–330 (1984).

W. Köhler, *Gestalt Psychology*, Liveright, New York, 1947.

S. W. Kuffler and J. G. Nicholls, *From Neuron to Brain*, Sinauer, Sunderland, Mass., 1976.

E. H. Land and J. J. McCann, "Lightness and Retinex Theory," *J. Opt. Soc. Am.* **61,** 1–11 (1971).

R. D. Luce, *Response Times: Their Role in Inferring Elementary Mental Organization*, Oxford University Press, Oxford, 1986.

H. A. Mallot, W. von Seelen, and F. Giannakopoulos, "Neural Mapping and Space-Variant Image Processing," *Neural Networks,* **3,** (1990).

D. Marr, *Vision*, Freeman, San Francisco, 1982.

W. S. McCulloch, *Embodiments of Mind*, MIT Press, Cambridge, Mass., 1965.

G. J. Mitchison and S. P. McKee, "Interpolation in Stereoscopic Matching. *Nature* **315,** 402–404 (1985).

S. E. Palmer, "The Psychology of Perceptual Organization: A Transformational Approach," in J. Beck, B. Hope, and A. Rosenfeld, eds., *Human and Machine Vision*, Academic Press, New York, pp. 269–340, 1983.

S. E. Palmer, E. Rosch, and P. Chase, "Canonical Perspective and the Perception of Objects," in J. Long and A. Baddeley, eds., *Attention and Performance IX*, Erlbaum, Hillsdale, N.J., pp. 135–151, 1981.

G. F. Poggio and T. Poggio, "The Analysis of Stereopsis," *Ann. Rev. Neurosci.* **7,** 379–412 (1984).

T. Poggio and S. Edelman, "A Network That Learns to Recognize Three-Dimensional Objects," *Nature* **343,** 263–266 (1990).

T. Poggio and F. Girosi, "Regularization Algorithms for Learning That Are Equivalent to Multilayer Networks," *Science* **247,** 978–982 (1990).

T. Poggio and W. Reichardt, "Visual Control of Orientation Behavior in the Fly (Parts i and ii)," *Quart. Rev. Biophys.* **3,** 311–439 (1976).

T. Poggio, V. Torre, and C. Koch, "Computational Vision and Regularization Theory," *Nature* **317,** 314–319 (1985).

M. Potter, "Meaning in Visual Search," *Science* **187,** 965–966 (1975).

Z. Pylyshyn, *Computation and Cognition*, MIT Press, Cambridge, Mass., 1985.

P. Quinlan, "Visual Object Recognition Reconsidered," *Behav. Brain Sciences* (in press).

V. S. Ramachandran, "Perception of Shape from Shading," *Nature* **331,** 163–166 (1988).

I. Rock, *Perception*, Scientific American Books, New York, 1984.

I. Rock and J. DiVita, "A Case of Viewer-Centered Object Perception," *Cogn. Psychol.* **19,** 280–293 (1987).

A. Rojer and E. L. Schwartz, "A Multiple-Map Model for Pattern Classification," *Neural Comput.* **1,** 104–115 (1989).

E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic Objects in Natural Categories," *Cogn. Psychol.* **8,** 382–439 (1976).

A. Rosenfeld, "Recognizing Unexpected Objects: A Proposed Approach," *Int. J. Pattern Recogn. Artif. Intell.* **1,** 71–84 (1987).

R. N. Shepard, "Toward a Universal Law of Generalization for Psychological Science," *Science* **237,** 1317–1323 (1987).

R. N. Shepard and L. A. Cooper, *Mental Images and Their Transformations*, MIT Press, Cambridge, Mass., 1982.

A. Sloman, "What Are the Purposes of Vision?," Technical Report No. CSRP 066, University of Sussex, U.K., 1987.

E. S. Spelke, "Origins of Visual Knowledge," in D. N. Osherson, S. M. Kosslyn, and J. M. Hollerbach, eds., *Visual Cognition and Action*, Vol. 2, MIT Press, Cambridge, Mass., pp. 99–128, 1990.

G. R. Stoner, T. D. Albright, and V. S. Ramachandran, "Transparency and Coherence in Human Motion Perception," *Nature* **344,** 153–155 (1990).

G. Stratton, "Vision without Inversion of the Retinal Image," in W. N. Dember, ed., *Visual Perception: The Nineteenth Century*, Wiley, New York, pp. 143–154, 1897/1964.

M. Tarr and S. Pinker, "Mental Rotation and Orientation-Dependence in Shape Recognition," *Cogn. Psychol.* **21,** 233–282 (1989).

J. T. Todd and E. Mingolla, "Perception of Surface Curvature and Direction of Illumination from Patterns of Shading," *J. Exp. Psychol. Human Perception and Performance* **9,** 583–595 (1983).

A. Treisman and G. Gelade, "A Feature Integration Theory of Attention," *Cogn. Psychol.* **12,** 97–136 (1980).

S. Ullman, "Filling in the Gaps: The Shape of Subjective Contours and a Model for Their Generation," *Biol. Cybernet.* **25,** 1–6 (1976).

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

S. Ullman, "Aligning Pictorial Descriptions: An Approach to Object Recognition," *Cognition* **32,** 193–254 (1989).

S. Ullman and R. Basri, "Recognition by Linear Combinations of Models," A.I. Memo No. 1152, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1990.

R. von der Heydt, E. Peterhans, and G. Baumgartner, "Illusory Contours and Cortical Neurons' Responses," *Science* **224,** 1260–1262 (1984).

H. von Helmholtz, "Unconscious Conclusions," in W. N. Dember, ed., *Visual Perception: The Nineteenth Century*, Wiley, New York, pp. 163–170, 1856/1964.

R. D. Walk, "Perceptual Learning," in E. C. Carterette and M. P. Friedman, eds., *Handbook of Perception*, Vol. IX, Academic Press, New York, pp. 257–298, 1978.

H. Wallach and D. N. O'Connell, "The Kinetic Depth Effect," *J. Exp. Psychol.* **45,** 205–217 (1953).

D. Weinshall, "Perception of Multiple Transparent Planes in Stereo Vision," *Nature* **341,** 737–739 (1989).

N. Weisstein and C. S. Harris, "Visual Detection of Line Segments: An Object-Superiority Effect," *Science* **186,** 752–755 (1974).

S. Zeki and S. Shipp, "The Functional Logic of Cortical Connections," *Nature* **335,** 311–317 (1988).

SHIMON EDELMAN
Weizmann Institute of Science

VISUAL RECOVERY

Neither objects nor properties of objects (such as shape, color, etc) exist inside our brains as such. When we see, computations are performed inside our heads which generate hypotheses about objects and their properties. To understand vision, the methods used to derive such perceptual hypotheses from visual images must be discovered.

Much of the research on computational vision over the past 35 years has concentrated on specific visual tasks, and has been concerned with how one can recover visual quantities necessary for carrying out such visual tasks as identifying and locating a known object so that a robot arm can grasp it. This research is sometimes referred to as belonging to the "recognition" school, since it deals with specific types of objects which must be identified and located. On the other hand, the "recovery" school has concentrated on the study of general visual capabilities, such as the ability to understand the shapes of general objects based on the distribution of surface markings (texture). This article reviews general visual recovery research and discusses how it relates to the recognition point of view.

Visual Problems

A very large variety of problems related to the interaction of autonomous mechanisms with their environments can potentially be solved using visual input. However, two classes of problems are commonly held to be touchstones for practical vision systems: successful navigation in a complex environment using visual information, and recognition of classes of common objects (such as people or trees) in a complex scene. A large proportion of the published papers on computer vision address, explicitly or implicitly, one or the other of these goals. If both were achieved, automatic systems would have many of the capabilities of the human visual system; but it is clear that constructing such systems presents great difficulties. These difficulties were realized during the 60s and the early 70s, after the failure of early attempts to build entire vision systems, ie, systems that exhibited some vertical integration and used knowledge at all levels, including domain-specific information. "In order to complete the construction of such systems, it is almost inevitable that corners be cut and many overly simplified assumptions be made" (Brady 1982). This results in a system capable of carrying out a limited number of tasks, but not enhancing our general understanding of vision. At about that time the recovery school of thought started to develop through the work of Marr (1982) and his colleagues. This school held that the majority of visual problems can be reduced to the following general problem: *from one or more images of a scene, derive an accurate three-dimensional description of the objects in the scene and quantitatively recover their properties (or at least those properties relevant to a given task).* If we can recover (reconstruct) an accurate description of our environment, we can navigate, avoid obstacles, and find specific locations. If we can recover the properties of an object (shape, reflectance, color, etc) we can use them to recognize (classify) the object. Thus, the recovery school of thought in computer vision emphasized the study of visual abilities, independently of a task.

Methodology of Visual Recovery

Even if we accept that the solution to vision problems lies in the recovery of the scene, it is not obvious how to proceed. Luckily, there is a standard way to design large, complex information systems, as research in computational fields has shown (Feldman, 1985). The system is divided into functional components or subsystems which break the overall task into autonomous parts. These subsystems are analyzed and the representations of information that they use and the language of communication among them are chosen. The subsystems are then tested individually, in pairs, and all together.

This approach can be used in building a visual system, using functionally autonomous subsystems that recover specific properties of the world from images. These subsystems are called modules. Visual recovery research is devoted to the study of such modules and their integration.

There is considerable evidence for the existence of such modules in the human visual system. One source of such evidence is the study of patients with visual disabilities that result from brain lesions. In addition, psychophysicists perform experiments in which a particular module of the human visual system is seemingly isolated, for example, Julesz's (1971) work on stereoscopic fusion without monocular cues, Land's (Land and McCann, 1971) work on the computation of lightness, Gibson's (1950) work on the perception of shape from texture, etc. Thus it seems that cues such as shading (image intensity variations), texture (distribution of surface markings), contours (image discontinuities), color, motion and stereo are very helpful in recovering properties of the three-dimensional (3-D) world from images.

Marr (1982) pointed out that perceptual processes, (ie, processes underlying visual abilities), must be understood at three levels:

1. The level of *computational theory*. We must develop, through rigorous mathematical treatment, the relationship between the quantity to be computed and the observations (data = image(s)). After this computational theory is developed, we can understand whether the given problem has a unique solution.
2. The level of *algorithms and data structures*. After the computational theory has been completed, we must design algorithms and data structures that, when applied to the input (image(s)), will output the desired quantity.
3. The level of *implementation*. After the two previous levels have been developed, we must implement the algorithm in hardware (serial or parallel).

If these three levels are fully understood, then we can say that we understand the perceptual process.

What Do We Want to Recover?

What should we attempt to recover from images in order to be able to accomplish visual tasks? The answer defines the nature of research on the theory of computer vision; that is, image understanding (IU) research which is not directed towards specific applications.

It is clear that one quantity that should be recovered from images is the shapes of objects. A large amount of visual recovery research is devoted to determining the shapes of imaged objects from image cues, such as shading, texture, contour, multiple views, motion, etc. If we can recover the geometry of the environment we can per-

form navigational tasks such as avoiding obstacles, finding passages, etc. In addition, if we can find an appropriate representation, we can use shape information for object recognition.

Shape is not the only thing we may want to recover; for example, if we can recover the three-dimensional velocity of a moving object we can catch it, avoid it, track it, etc. Or, we may want to determine the velocity with which every image point moves, in the case of images obtained by a moving sensor (optic flow). We may also want to recover the colors of objects; to recover then pose (spatial position and orientation) of a known object; to determine the discontinuities of the image intensity that correspond to physical discontinuities; to determine a segmentation of the image that corresponds to some well-defined segmentation of the scene; or we may want to recover (or restore) the ideal image from the actual image, which is corrupted by noise, etc. Evidently, it would not be possible to review all of these topics in this paper. Most of the articles in this encyclopedia that deal with vision are devoted to general or specific recovery problems. In this article, attention will focus on the problem of recovering shape.

Since this article deals with shape recovery, shape will be defined, several commonly used shape representations will be introduced, and the different kinds of projections (imaging geometries) used in the literature will be reviewed. The modules of shape from shading, texture, contour, stereo and motion will be studied, in brief: "shape from x". The constraints relating 3-D shape to observable image data, as well as algorithms that aim to reconstruct the 3-D world from single monocular cues, will be described. After discussing the limitations of these approaches, there will be discussions on how to combine multiple cues, and an outline of the foundations for the active vision paradigm, which provides the basis for the unique and robust computation of 3-D shape, will be included.

GEOMETRIC CORRESPONDENCE BETWEEN SCENE AND IMAGE

Different imaging projections have different properties that influence the design of shape recovery modules. Described here are the most commonly used projections.

Perspective Projection

Consider an ideal pinhole at a fixed distance in front of an image plane (Fig. 1). Assume that only light coming through the pinhole can reach the image plane. Given that light travels along straight lines, each point in the image corresponds to a particular direction defined by the ray from that point through the pinhole. This is perspective projection. In the sequel, in order to simplify the equations, the nodal point of the eye (the pinhole) will be regarded as being behind the image plane, as shown in Figure 2. The optical axis is defined to be the perpendicular from the pinhole to the image plane. A Cartesian coordinate system is introduced with the origin at the nodal point and the z-axis aligned with the optical axis and pointing toward the image. Let A be any point in front of the camera. Assume that nothing lies on the ray from point A to the nodal point O. Compute the position of the image A' of A in the image plane. Let $V = (X,Y,Z)$ be the vector connecting O to A and $V' = (x,y,f)$ be the vector connecting O to A', with f the focal length, ie, the distance of the image plane from the nodal point O. Then (x,y) are the coordinates of A' on the image plane in the naturally induced coordinate system with origin the point of the intersection of the image plane with the optical axis, and axes x and y parallel to the axes of the camera coordinate system OX and OY. It is trivial to see that

$$x = \frac{fX}{Z}, y = \frac{fY}{Z} \tag{2.1}$$

Equations 2.1 relate the world coordinates of a point to the image plane coordinates of its image. Very often, to further simplify the equations, we assume $f = 1$ without loss of generality.

Orthographic Projection

If, in the perspective projection model, we have a scene plane that lies parallel to the image plane at $Z = Z_0$, then we define the magnification, μ, as the ratio of the distance between two points measured in the image to the distance between the corresponding points in the scene plane. So, for a small interval (dX, dY, O) on the scene plane and the corresponding small interval (dx, dy) in the image, we have

$$\mu = \frac{(dx)^2 + (dy)^2}{(dX)^2 + (dY)^2} = \frac{f}{Z_0}$$

Thus a small object in the scene at average distance Z_0 will produce an image that is magnified by μ. Evidently

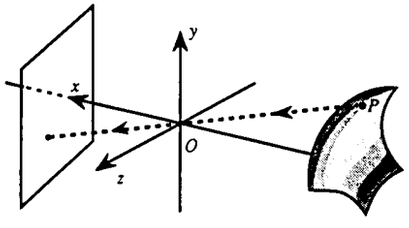


Figure 1. Perspective projection (image plane in the back).

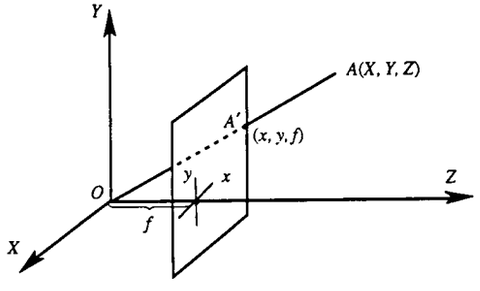


Figure 2. Perspective projection (image plane in front).

the magnification is approximately constant when the depth range of the scene is small relative to the average distance of the scene points from the camera. In this case, equations 2.1 become

$$x = \mu X, y = \mu Y \tag{2.2}$$

with $\mu = f/Z_0$ and Z_0 the average value of the depth Z. For convenience, if $\mu = 1$, equations 2.2 further simplify to

$$x = X, y = Y \tag{2.3}$$

Equations 2.3 define the orthographic projection model, where the rays are parallel to the optical axis (Fig. 3). The difference between orthography and perspective is small when the distance to the scene is much larger than the variation in distance among objects in the scene. A rough rule of thumb is that perspective effects are significant when a wide angle lens is used, while images taken by telephoto lenses tend to approximate orthographic projection; of course, this is not exact (Horn, 1986).

Paraperspective Projection

Orthographic projection is a very rough approximation of the projection of light on the fovea, but it is unrealistic for many machine vision applications. Perspective projection, on the other hand, involves more complicated equations and makes the analysis of some problems difficult. Paraperspective projection is a good approximation of perspective; it lies between orthography and perspective. A version of paraperspective projection was first introduced by Ohta and co-workers (1981). Let a coordinate system $OXYZ$ be fixed with respect to the camera, with Z axis pointing along the optical axis and O the nodal point of the eye. Again, we consider the image plane perpendicular to the X axis at the point (0,0,1) (ie, the focal length $f = 1$, without loss of generality). Consider a small planar surface patch SP having the equation $Z = pX + qY + C$ (Fig. 4). Under perspective, any point $(X,Y,Z) \in SP$ is projected onto the point $(X/Z, Y/Z)$ on the image plane. Now consider the plane $Z = d$, where d is the Z-coordinate of the centroid C of SP. Paraperspective projection involves two steps:

1. SP is projected onto $Z = d$. This projection is performed using the rays that are parallel to the central projecting ray OC.
2. The projection of SP on $Z = d$ is projected perspectively onto the image plane. Since $Z = d$ is parallel to the image plane, this projection is a magnification by a factor $1/d$.

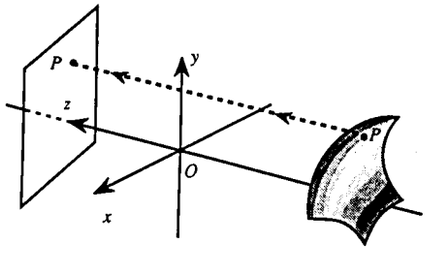


Figure 3. Orthographic projection.

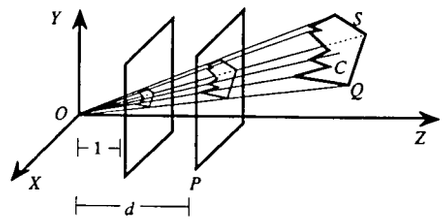


Figure 4. Paraperspective projection. Plane P is put in front of the surface S for pictorial clarity.

Figure 5 illustrates a cross sectional view of the projection process sliced by a plane perpendicular to the XZ plane and which includes the central projecting ray. Paraperspective decomposes the projection of the scene onto the image plane into two parts. Step (a) incorporates the foreshortening distortion and part of the position effect, and step (b) incorporates both the distance and the position effects.

Paraperspective projection has nice mathematical properties, since it is an affine transformation. It has been successfully used in many areas of computer vision, such as shape from texture, shape from contour, object recognition, and the like. See Aloimonos (1990b) for applications of paraperspective projection, as well as other perspective approximations.

WHAT DO WE MEAN BY SHAPE?

A visual system analyzes images and produces descriptions of what is imaged. A description might include information about the shapes of the objects in the scene, but the shape of an object does not have a unique description; one can think of descriptions at many levels of detail and from many points of view. As Horn (1986) suggests, "we can avoid this potential philosophical snare by considering the task for which the description is intended. We don't want just any description of what is imaged, but one that allows us to take appropriate action." A reasonable first approximation to describing the shape of an object is to represent the local orientation of its surface. Only this level of description will be considered here. Global shape representations can also be used, for example the one based on superquadrics (Pentland, 1986; Bajcsy and Solina, 1987); they are appealing because they represent complex shapes using only a few numbers, but the inverse problem (finding this description from an image) still

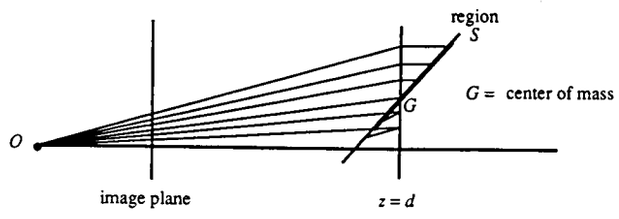


Figure 5. Cross-sectional view of paraperspective.

needs to be addressed. Various qualitative shape descriptors have been developed in the two-dimensional literature (Pavlidis, 1980), and a few in the 3-D case (Mumford, 1987), for use in object recognition; but these descriptors do not provide solutions to the general recovery task. Finding robust shape descriptions is an open research problem that will probably require advanced mathematical tools for its solution, and therefore this section is confined to the local descriptions of shape based on surface orientation.

Surface orientation is usually represented by the orientation of the surface normal vector. In the following subsections it is shown how the shape of a visible surface can be reconstructed from local orientation information.

Surface Orientation and Shape

The normal vector $\bar{n}$ to the surface $Z = Z(X,Y)$ at the point (X,Y,Z) is

$$\bar{n} = \left(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, -1\right) \Big/ \left[\left(\frac{\partial Z}{\partial X}\right)^2 + \left(\frac{\partial Z}{\partial Y}\right)^2 + 1\right]^{1/2}$$

Let $(x = fX/Z, y = fY/Z)$ be the image of (X,Y,Z). If (dx, dy) is a small displacement in the image, corresponding to a small displacement (dX, dY, dZ) on the surface, then

$$dX = \frac{dx \cdot Z + x\, dZ}{f}, \quad dY = \frac{dy \cdot Z + y\, dZ}{f}$$

Given that $Z(X + dX, Y + dY) = Z(x + dx, y + dy)$, if we expand both sides of this equation in a Taylor series and ignore the higher order terms, we get

$$\frac{\partial Z}{\partial X}\frac{Z}{f - x\frac{\partial Z}{\partial X} - y\frac{\partial Z}{\partial Y}}dx + \frac{\partial Z}{\partial Y}\frac{Z}{f - x\frac{\partial Z}{\partial X} - y\frac{\partial Z}{\partial Y}}dy = \frac{\partial Z}{\partial x}dx + \frac{\partial Z}{\partial y}dy$$

so that

$$\frac{\partial Z}{\partial x} = \frac{Z\frac{\partial Z}{\partial X}}{f - x\frac{\partial Z}{\partial X} - y\frac{\partial Z}{\partial Y}} \text{ and } \frac{\partial Z}{\partial y} = \frac{Z\frac{\partial Z}{\partial Y}}{f - x\frac{\partial Z}{\partial X} - y\frac{\partial Z}{\partial Y}}$$

From these equations we see that if $\partial Z/\partial X$, $\partial Z/\partial Y$ are known, the quantity

$$\frac{Z(x + dx, y + dy)}{Z(x,y)}$$

can be computed. This means that if the surface normal is known as a function of position (x,y) in the image, then the depth function $Z(x,y)$ can be computed up to a constant factor. The constant is undetermined; the surface can be small and near the camera or large and far away.

Surface Orientation and Shape Under Orthography

Under orthographic projection, the image coordinates of a point are equal to the corresponding scene coordinates, ie, $(x,y) = (X,Y)$. So

$$\left(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}\right) = \left(\frac{\partial Z}{\partial x}, \frac{\partial Z}{\partial y}\right).$$

Since

$$Z(x + dx, y + dy) - Z(x,y) = \frac{\partial Z}{\partial x}dx + \frac{\partial Z}{\partial y}dy + (\text{higher order terms}),$$

we see that $Z(x,y)$ can be computed up to a constant additive term. Thus, if we know the surface orientation under orthography, we know the surface shape, but we do not know its distance.

Other Coordinate Systems

Let $p = \partial Z/\partial X$, $q = \partial Z/\partial Y$ at the point of the surface $Z = Z(X,Y)$. We have seen that the surface normal vector is

$$\frac{(p,q,-1)}{(p^2 + q^2 + 1)^{1/2}}$$

The coordinates

$$(a,b,c) = \left(\frac{p}{k}, \frac{q}{k}, \frac{-1}{k}\right) \text{ with } k = (p^2 + q^2 + 1)^{1/2}$$

define the position of a point on the *Gaussian sphere*. This position can also be defined in terms of latitude and longitude angles. Another commonly used representation is in terms of slant and tilt, (σ, τ), where slant is the tangent of the latitude angle and tilt is the longitude angle. It is easy to see that

$$\sigma = \cos^{-1}\left(\frac{1}{\sqrt{1 + p^2 + q^2}}\right)$$

$$\tau = \tan^{-1}\left(\frac{q}{p}\right)$$

The parameterization of the local surface normal that uses the partial derivatives $p = \partial z/\partial x$, $q = \partial z/\partial y$, gives rise to the concept of gradient space (see, eg, Shafer and coworkers, 1983). The parameterization has the disadvantage that the partial derivatives can become infinite at occluding boundaries, ie, at places where the surface turns away from the viewer; a similar problem arises with the slant-tilt representation. Ikeuchi and Horn (1981) therefore used a different parameterization (f,g) of surface orientation, which they called stereographic space. f and g are related to p and q by

$$f(g) = \frac{2p(q)[\sqrt{1 + p^2 + q^2} - 1}{p^2 + q^2}$$

Using the Gaussian sphere formalism, they showed that gradient space corresponds to projecting the Gaussian sphere from its center onto a plane tangent to the sphere at its north pole, whereas stereographic space corresponds to projecting from the south pole (see Figs. 6 and 7).

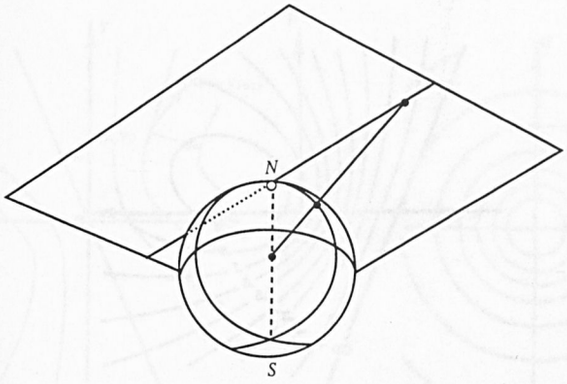


Figure 6. Projection of Gaussian sphere on gradient space. (Reproduced from Horn (1985).)

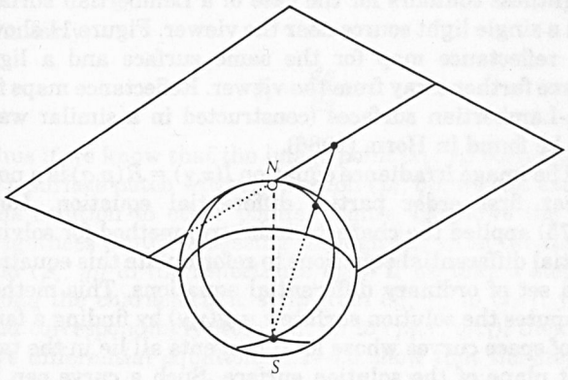


Figure 7. Stereographic projection. (Reproduced from Horn (1985).)

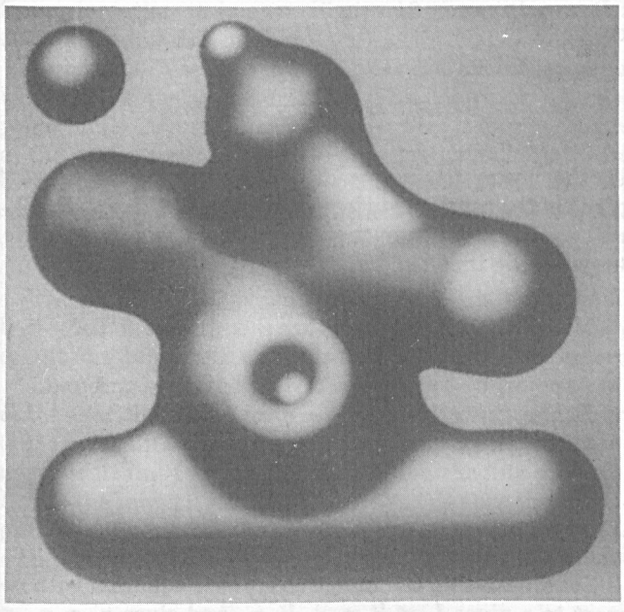


Figure 8. Shaded surface.

For a clear discussion of different shape representations, see Horn, (1986).

SHAPE FROM x

Modules that recover surface orientation from various cues in the image are called *shape from x* modules. Some of these modules operate directly on the image, while others operate on some intermediate representation created from the image. Shape from shading falls in the first category while shape from texture and contour fall in the second. As regards shape from stereo and shape from motion, some researchers put them in the first category while others put them in the second. The recovery problems defined by these modules are usually ill posed, so additional constraints must be introduced.

Shape from Shading

The recovery of surface orientation from gray level variations (shape from shading) was first studied by Horn and his colleagues at MIT. The analysis was done under orthography. Figure 8 shows an image that contains shading. Humans can easily perceive, at least qualitatively, the shape of the imaged surface. In this section we describe methods of recovering surface orientation from shading, together with other assumptions to be described later.

In general, the amount of light reflected by a surface element (the surface radiance) depends on its microstructure, on its optical properties, and on the angular distribution and state of polarization of the incident illumination. For some surfaces, the fraction of incident illumination (irradiance) reflected in a particular direction depends only on the surface orientation. The reflectance of such a surface can be represented by a function $f(i,g,e)$ of the angles i = incident, g = phase and e = emergent, as they are defined in Figure 9. For example, in perfect specular (mirror-like) reflection, the incident angle equals the emergent angle and the incident, emergent and normal vectors lie in the same plane; the phase angle is given by $g = i + e$. Thus the reflectance function is

$$f(i,e,g) = \begin{cases} 1 \text{ if } i = e \text{ and } i + e = g \\ 0 \text{ otherwise} \end{cases}$$

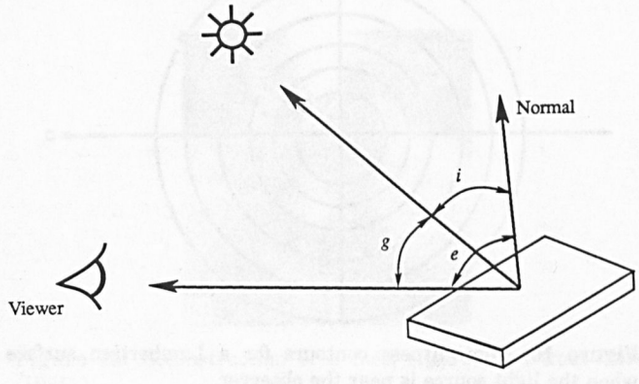


Figure 9. Geometry of reflection.

The most widely used model of surface reflectance is defined by the function $f(i,e,g) = \rho \cos i$, where ρ is constant for a given surface and is called the albedo constant. This function defines the reflectance of a perfectly diffuse (Lambertian) surface which appears equally bright from all viewing directions; the cosine of the incident angle compensates for the foreshortening of the surface as seen from the light source.

In orthographic projection, the viewing direction and hence the phase angle g are constant for all surface elements. So, for a fixed light source and viewer geometry and a given surface material, the ratio of radiance to irradiance depends only on the surface normal vector. Furthermore, suppose that each surface element receives the same irradiance. Then the surface radiance, and hence the image intensity $I(x, y)$, depends only on the surface normal vector.

When expressed in terms of the surface normal coordinates $p = \partial z/\partial x$, $q = \partial z/\partial y$, the reflectance function is called the *reflectance map* and is denoted by $R(p,q)$. This map provides a uniform representation for a given surface material for a particular light source, surface normal, and viewer geometry. A comprehensive discussion of reflectance maps for a variety of surface and light source conditions has been given by Horn (1977). A unified approach to the specification of reflectance maps has been given in Horn and Sjoberg, (1979).

Under orthographic projection, expressions for $\cos i$, $\cos e$, and $\cos g$ can be easily derived from the surface normal vector $(p, q, -1)$, the light source vector $(p_s, q_s, -1)$, and the vector $(0, 0, -1)$ that points in the direction of the viewer. For a Lambertian reflectance function these expressions give

$$R(p,q) = \frac{\rho(1 + p\,p_s + q\,q_s)}{\sqrt{(1 + p^2 + q^2)}\sqrt{(1 + p_s^2 + q_s^2)}}$$

where ρ is the albedo constant. Under perspective projection, the expressions are not known exactly, but recent results indicate that they are similar (Shafer and co-workers, 1983).

Using fixed light sources and fixed reflectance characteristics, the reflectance map associates a brightness value with each surface orientation. Figure 10 shows isobrightness contours for the case of a Lambertian surface and a single light source near the viewer. Figure 11 shows the reflectance map for the same surface and a light source farther away from the viewer. Reflectance maps for non-Lambertian surfaces (constructed in a similar way) can be found in Horn, (1986).

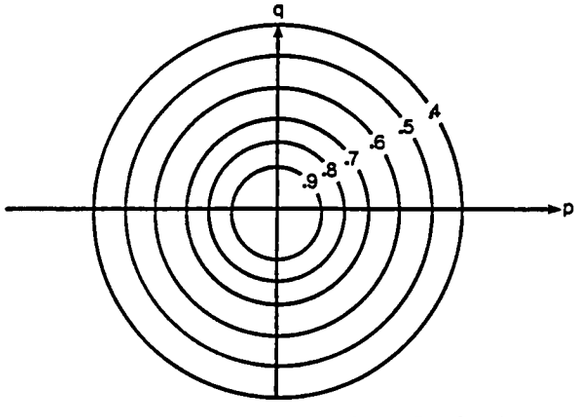


Figure 10. Isobrightness contours for a Lambertian surface when the light source is near the observer.

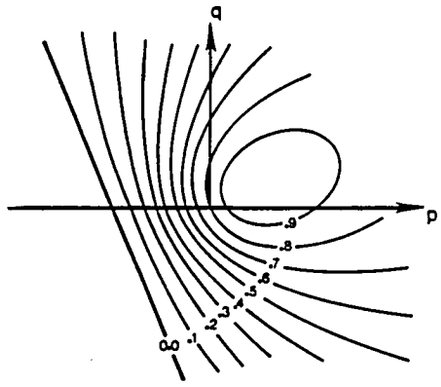


Figure 11. Isobrightness contours for a Lambertian surface when the light source is removed from the observer. (Reproduced from Horn (1977).)

The image irradiance equation $I(x,y) = R(p,q)$ is a nonlinear first order partial differential equation. Horn (1975) applied the characteristic strip method for solving partial differential equations to reformulate this equation as a set of ordinary differential equations. This method computes the solution surface $z = g(x,y)$ by finding a family of space curves whose local tangents all lie in the tangent plane of the solution surface. Such a curve can be specified by a one-parameter family of points $(x(s), y(s), z(s))$, where s is the distance along the curve.

Differentiating with respect to s gives

$$p\frac{dx}{ds} + q\frac{dy}{ds} - \frac{dz}{ds} = 0$$

or

$$(p,q,-1) \cdot \left(\frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds}\right) = 0,$$

ie, the vector $(dx/ds, dy/ds, dz/ds)$ lies in the tangent plane of the solution surface. Trivially, the vector $(R_p, R_q, pR_p + qR_q)$ also lies in that plane. From this observation, we conclude that

$$\frac{dx}{ds} = R_p \tag{4.1}$$

$$\frac{dy}{ds} = R_q \tag{4.2}$$

$$\frac{dz}{ds} = pR_p + qR_q \tag{4.3}$$

where the subscripts denote partial differentiation.

Differentiating the image irradiance equation with respect to x gives $I_x = R_p p_x + R_q q_x$, and since $p_y = g_{xy} =$

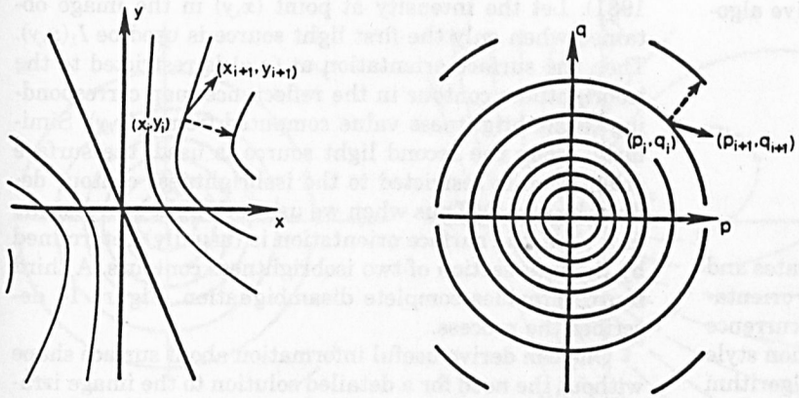


Figure 12. Isobrightness contours $(x_i, y_i) \rightarrow (p_i, q_i)$. (Reproduced from Brady (1982).)

$g_{yx} = q_x$ we have $I_x = R_p p_x + R_q p_y$, and consequently

$$I_x = \frac{dp}{ds} \tag{4.4}$$

Similarly,

$$I_y = \frac{dq}{ds} \tag{4.5}$$

Thus if we know that the image point (x_k, y_k) corresponds to a surface patch with orientation (p_i, q_i), we can extend this solution to other points. Figure 12 shows the isobrightness contours passing through (x_i, y_i) in the image and (p_i, q_i) in the reflectance map. If we take a step ds along the characteristic strip from (x_i, y_i) to (x_{i+1}, y_{i+1}), and correspondingly from (p_i, q_i) to (p_{i+1}, q_{i+1}), then the five differential equations 4.1–4.5 show that the step in the image is in direction (R_p, R_q), ie, along the normal to the isobrightness contour in the reflectance map. In the same way, the step in the reflectance map is in the direction normal to the isobrightness contour computed in the image. Thus, if we know the reflectance map we can compute the surface orientations at a sequence of points along a characteristic strip starting from a point where the surface orientation is known. Figure 13 shows the results obtained using this method.

In order to use this method we need an initial point with known surface orientation. The algorithm also depends on the assumption that the surface is locally convex at the initial point. At this stage, researchers began to be concerned about conditions under which the method works, as well as about uniqueness issues. These questions were important in subsequent research on recovery (Barrow and Tenenbaum, 1981a).

The problem is ill-posed so additional constraints will be needed. A smoothness constraint, along with boundary conditions, provides a unique solution, as described below.

Bounding or occluding contours provide boundary conditions for the shape from x problem. Ikeuchi and Horn (1981) used these conditions in conjunction with a smoothness constraint to solve the shape from shading problem. If I_{ij} is the intensity at point (i,j), and (f,g) are the stereographic coordinates of the surface orientation, we look for a surface (f_{ij}, g_{ij}), $(i,j) \in$ image that minimizes

$$e = \sum_i \sum_j (s_{ij} + \lambda r_{ij}),$$

where

$$s_{ij} = \frac{1}{4}[(f_{i+1,j} - f_{ij})^2 + (f_{i,j+1} - f_{ij})^2 + (g_{i+1,j} - g_{ij})^2 + (g_{i,j+1} - g_{ij})^2]$$

and

$$r_{ij} = (I_{ij} - R(f_{ij}, g_{ij}))$$

The first term in the sum represents departure from smoothness while the second represents departure from the constraint defined by the image irradiance equation. Thus the surface that minimizes e best satisfies the image irradiance equation and is also as smooth as possible. The parameter λ defines the relative importance of the smoothness and the irradiance constraint. We minimize e by differentiating with respect to f_{ij}, g_{ij} and setting the resulting derivatives equal to zero. This gives the follow-

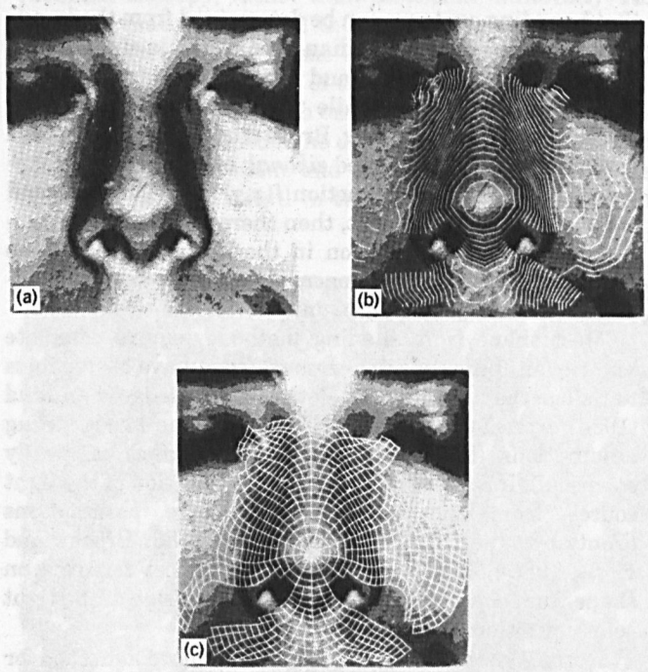


Figure 13. Reconstruction of a face. (Reproduced from Horn (1985).)

ing recurrence relations as the basis of an iterative algorithm (Ikeuchi and Horn, 1981):

$$f_{ij}^{(n+1)} = \overline{f_{ij}^{(n)}} + \lambda(I_{ij} - R(f_{ij}^{(n)}, g_{ij}^{(n)})) \frac{\partial R}{\partial f}$$

$$g_{ij}^{(n+1)} = \overline{g_{ij}^{(n)}} + \lambda(I_{ij} - R(f_{ij}^{(n)}, g_{ij}^{(n)})) \frac{\partial R}{\partial g}$$

where the superscripts in parentheses denote iterates and the bars denote local averages. Since the surface orientation at the occluding boundaries is known, this recurrence propagates information inwards and in a relaxation style computes the orientation everywhere. This algorithm works well for many images, but there is no proof that it converges. An important aspect of the algorithm is graceful degradation under errors in the placement of the light source, the surface orientation on the boundary, and the nature of the surface reflectivity. The algorithm also does not guarantee integrability of the resulting surface orientation function. Horn and Brooks (1986) attempted to remedy this deficiency; Frankot and Chellappa (1987) developed a method of enforcing integrability.

Other authors proposed smoothness constraints derived from the fact that the integral of depth around a closed path in the image is zero (Brooks, 1979; Strat, 1979). Woodham observed that the shape from shading problem can be solved uniquely if a global assumption is made about the shape of the surface, for example that it is convex, a ruled surface, or a generalized cylinder (1981).

The mathematical properties of the image irradiance equation were studied by Bruss (1980). She showed that a continuous image irradiance equation can have discontinuous solutions, and that the curvature of a surface cannot be identified in general from its image. However, Bruss proved that there is only one solution which is convex, and that bounding contours can be determined from the image only when the image irradiance equation is singular, ie, the reflectance function R and its first-order partial derivatives are continuous, while the intensity function I is discontinuous in x and/or y. Bruss studied singular image irradiance equations, called *eikonal*, of the form $p^2 + q^2 = I(x,y)$. If the intensity function $I(x,y)$ vanishes to second order at the singular point, then there is exactly one positive locally convex solution in the neighborhood of the singular point. In consequence, if there is a closed bounding contour, the solution is unique.

Most shape from shading methods require complete knowledge of the reflectance map. There have been efforts to reduce the need for such detailed knowledge. Pentland (1984) extracts information locally, but he needs strong assumptions (for example, that the surface is locally spherical). It is possible to recover the position of the light source from the image under some assumptions (Pentland, 1982; Lee and Rosenfeld, 1985; Brooks and Horn, 1985). This is important, since most research on shape from shading assumes exact knowledge of the light source position.

Horn, Woodham and Silverman developed a method for computing shape from shading using multiple (known) light sources; it is called *photometric stereo* (Woodham, 1981). Let the intensity at point (x,y) in the image obtained when only the first light source is used be $I_1(x,y)$. Then the surface orientation at (x,y) is restricted to the isobrightness contour in the reflectance map corresponding to the brightness value computed from $I_1(x,y)$. Similarly, when the second light source is used, the surface orientation is restricted to the isobrightness contour defined by $I_2(x,y)$. Thus when we use both light sources, one at a time, the surface orientation is (usually) determined by the intersection of two isobrightness contours. A third source provides complete disambiguation. Figure 14 describes the process.

One can derive useful information about surface shape without the need for a detailed solution to the image irradiance equation. For example, information from isobrightness contours might be beneficial (Koenderink and Van Doorn, 1980). The human visual system may obtain global shape information from shading without constructing a surface normal map.

Highlights in images of objects with specularly reflecting surfaces provide significant information about the surfaces. Coleman and Jain (1982) presented a method using four-source photometric stereo to identify and correct for specular reflection components. Blake (1985) assumes smooth surfaces and single point specularities and develops a computational theory for shape extraction based on the disparities of the specularities in a pair of images (specular stereo). Healey and Binford (1987) derived relationships between the properties of a specular feature in an image and local properties of the corresponding surface. Wolff (1986, 1987a, 1987b) studied shape extraction techniques from multiple images using spectral and polarization properties. Research has also begun on non-Lambertian surfaces, on illumination models due to sun and sky, etc. A comprehensive collection of papers on the topic is given in (Horn and Brooks, 1989).

Shape from Texture

Texture provides an important source of information about the orientations of surfaces. Figures 15 and 16 show the perspective images of some natural surfaces. It seems that a human can easily perceive the shapes of the surfaces. To recover shape from texture, the distorting effects of the surface orientation and the imaging geometry must be distinguished from the properties of the texture on which the distortion acts. This requires that assumptions be made about the texture. The problem of recovering the orientation of a planar surface from texture for the case of planes has been extensively studied; see (Gibson, 1950; Witkin, 1981; Stevens, 1981; Bajcsy and Lieberman, 1976; Kender, 1979, 1980; Kanatani, 1984; and Aloimonos, 1986). These studies were based on different assumptions about the texture and the imaging geometry.

The process of image formation (projection) introduces distortions into the appearance of the scene. In general, the distortions can be considered as due to two effects: the *distance* effect (objects appear larger when they are closer to the camera), and the *foreshortening* effect (the distortion depends on the angle between the surface normal and the line of sight). The orthographic projection model cap-

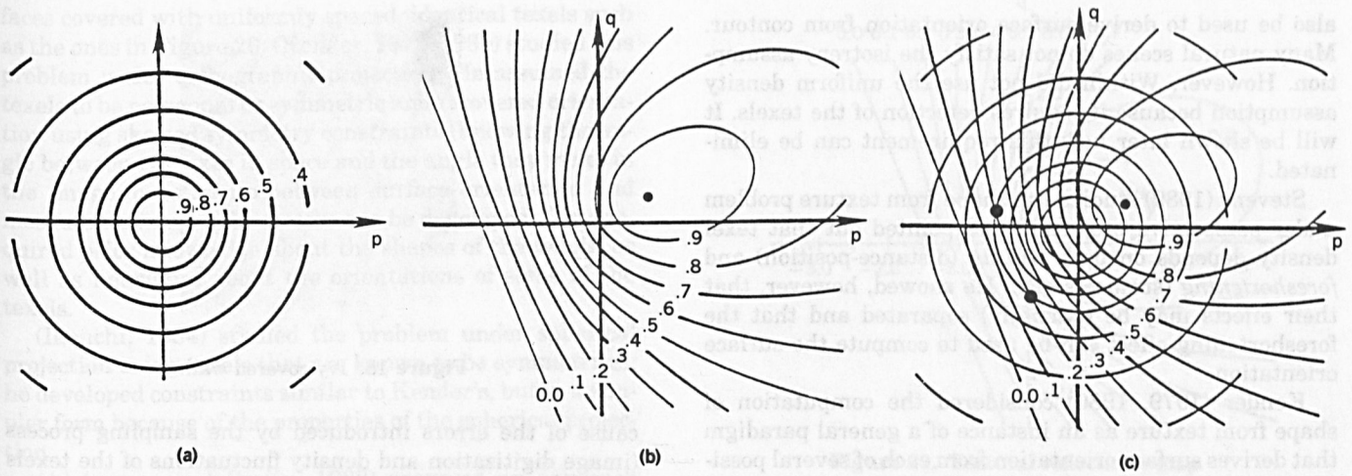


Figure 14. An illustration of photometric stereo. (Reproduced from Brady (1982).)

tures only the foreshortening effect and ignores the distance effect. Therefore, methods for shape from texture which use orthographic projection are valid only in a limited domain. The perspective projection model captures both effects, but the resulting algorithms involve the solution of nonlinear equations, and numerical errors limit their accuracy.

The first to approach the shape from texture problem was Gibson (1950). Trying to develop a theory of how humans perceive surface orientation from texture, he suggested that textures consist of small elements, which we shall call texels. Of course, these texels may be arranged very irregularly. We assume, however, that the texels are uniformly distributed on the scene plane, in the sense that each unit area on that plane contains approximately the same number of texels. In the image, however, the texel density may not be uniform; it may vary (linearly) with position. The gradient (magnitude and direction of maximum rate of change) of texture density in the image then determines the surface orientation; the magnitude depends on the surface slant, and the direction on the tilt.

Bajcsy and Lieberman (1976) used the two-dimensional Fourier power spectrum to detect the texture gradient. Their theory assumes that all the texture elements have the same size.

Witkin (1981) presented an approach that assumed directional isotropy, rather than positional uniformity. He assumed that the edges of the texels have uniformly distributed orientations. In the image, the orientations will be biased; the magnitude of the bias depends on the surface slant and its direction on the tilt. Based on an orthographic projection model, he derived maximum likelihood estimators for the slant and tilt. Witkin's work will be described in more detail in the next section since it can

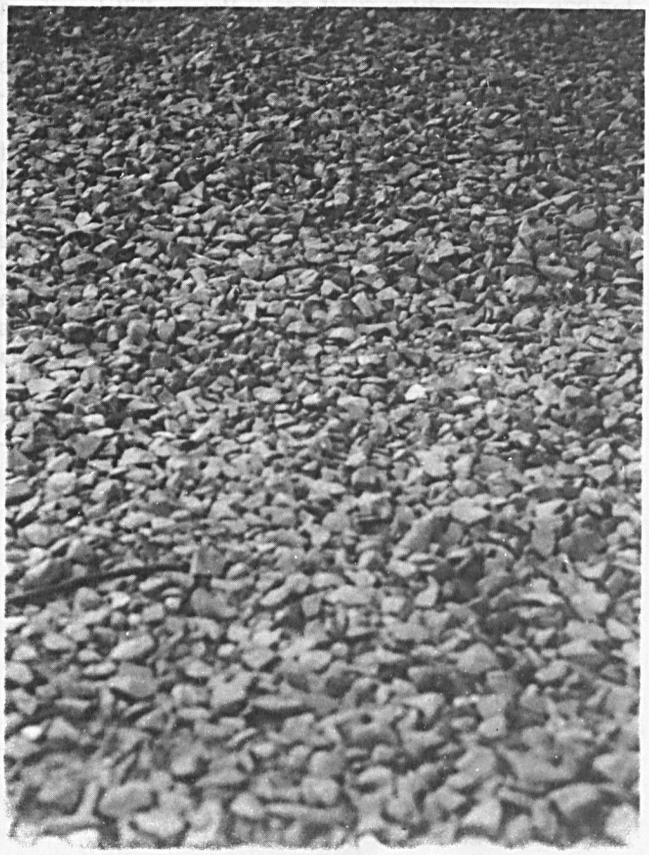


Figure 15. Textured surface (gravel).

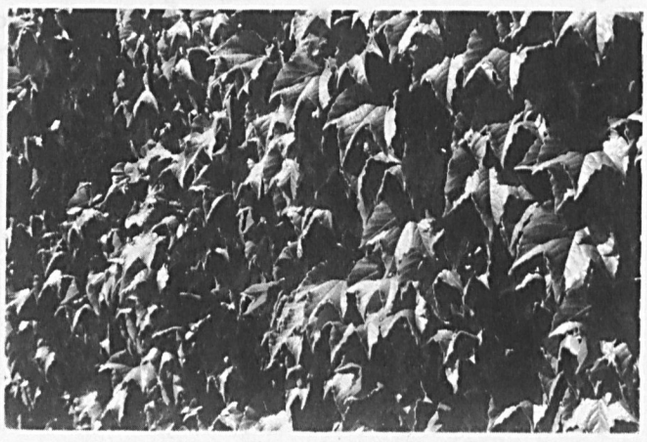


Figure 16. Textured surface (ivy).

also be used to derive surface orientation from contour. Many natural scenes do not satisfy the isotropy assumption. However, Witkin did not use the uniform density assumption because it requires detection of the texels. It will be shown later how this requirement can be eliminated.

Stevens (1980) studied the shape from texture problem under perspective projection and pointed out that texel density depends on both *scaling* (distance-position) and *foreshortening* (surface slant). He showed, however, that their effects may be (partially) separated and that the foreshortening effect can be used to compute the surface orientation.

Kender (1979, 1980) considered the computation of shape from texture as an instance of a general paradigm that derives surface orientation from each of several possible image observables. He assumes that texels are extracted from the image, and that each texel belongs to a planar surface. He defines a set of normalized texel property maps (NTPM) that generalize the reflectance map in shape from shading. If we assume that the texels all lie in a plane, and all have the same values of a given property (eg, diameter), we can derive constraints on the orientation of the plane.

Recent work (Aloimonos, 1988a) has developed a robust method of estimating the orientation of a planar surface based on the uniform density assumption. Let image regions R_1 and R_2 have areas S_1 and S_2 and contain k_1 and k_2 texels, respectively. Under paraperspective projection, the areas of the corresponding regions on the scene plane are $T_1 = S_1 c^2 \sqrt{a + p^2 + q^2}/(1 - A_1 p - B_1 q)^2$ and $T_2 = S_2 c^2 \sqrt{1 + p^2 + q^2}/(1 - A_2 p - B_2 q)^2$, respectively, where (A_1, B_1) and (A_2, B_2) are the centroids of R_1, R_2 and the scene plane has equation $Z = pX + qY + c$. By the uniform density assumption we have $k_1/T_1 = k_2/T_2$, and this can be transformed to give

$$\left[\left(\frac{k_2}{k_1}\frac{S_1}{S_2}\right)^{1/3} A_2 - A_1\right] p + \left[\left(\frac{k_2}{k_1}\frac{S_1}{S_2}\right)^{1/3} B_2 - B_1\right] q = \left(\frac{k_2}{k_1}\frac{S_1}{S_2}\right)^{1/3} - 1$$

This equation represents a line in p-q space; thus comparing the counts of texels in two image regions constrains (p,q) to lie on a line in gradient space. Ideally, using two pairs of image regions we can solve for p and q. But because of the errors introduced by the sampling process (image digitization and density fluctuations of the texels in the regions), this will give unreliable results. To obtain a robust result we consider many pairs of image regions. Each pair gives us a line in the gradient space, and the desired solution is the point whose sum of distances from all the lines is minimum (Fig. 17).

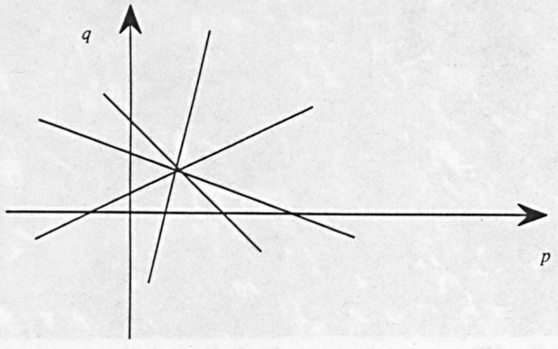


Figure 17. Intersection of lines in gradient space.

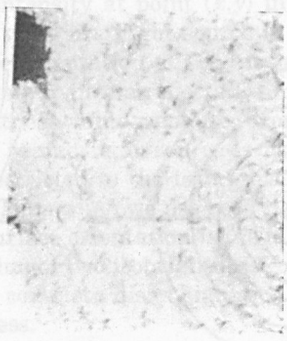


Figure 18. Ivy-covered wall.

The above method requires that the texels be identified so they can be counted. A more realistic approach uses the total length of edges in an image region; assuming that these are texel edges, their total length should be proportional to the number of texels. Using this method one can recover the orientations of planar surfaces in real-world scenes. For example, Figure 18 shows the image of an ivy-covered wall with orientation (slant = 20°, tilt = 0°). Figure 19 shows the extracted edges; this edge image was input to an algorithm, that using the modified uniform density assumption, recovered (slant = 24.5°, tilt = 5.6°). For other examples and a theoretical treatment, see Aloimonos, (1988a). For other work using a uniform density approach, see Kanatani and Chou, (1989).

Kanatani (1984) used the second Fourier harmonics of the number of intersections between texels and parallel scan lines to find planar surface orientation, under orthographic projection, under the assumption that the texture is directionally isotropic; for other uses of the isotropy assumption see the section on Shape from Contour. Ohta and Kanade (1985) separated the image texels into types and derived surface orientation information from the area ratios of pairs of texels of the same type.

Research on shape from texture for nonplanar surfaces has been restricted to idealized domains involving sur-

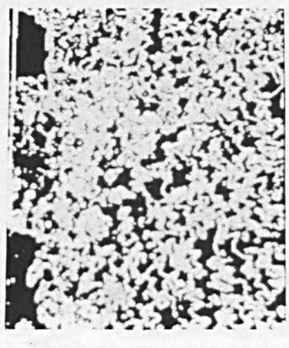


Figure 19. Edge image of the wall.

faces covered with uniformly spaced, identical texels such as the ones in Figure 20. (Kender, 1979, 1980) studied this problem under orthographic projection. He assumed the texels to be polygonal or symmetric and recovered orientation using skewed symmetry constraints (knowing the angle between two axes in space and the angle they make in the image, constraints between surface orientation and measurable image parameters can be developed). This required prior knowledge about the shapes of the texels, as well as heuristics about the orientations of some of the texels.

(Ikeuchi, 1984) studied the problem under spherical projection using texels that are known to be symmetrical; he developed constraints similar to Kender's, but in a simpler form because of the properties of the spherical projection.

(Aloimonos and Swain, 1988) proposed an approach that applies the methods used in shape from shading to the problem of shape from texture. Assume that all the texels are approximately planar and have the same area, and that we use paraperspective projection. Let S_I be the area of an image texel, S_W the area of the corresponding scene texel, (A,B) the centroid of the image texel, and d the range to the scene texel; then (assuming focal length = 1) it can be shown that

$$S_I = \frac{S_W}{d^2} \frac{1 - Ap - Bq}{\sqrt{1 + p^2 + q^2}}$$

where (p,q) is the gradient of the plane containing the scene texel. If we call S_I the "textural intensity," and S_W/d^2 the "textural albedo," the above equation is very similar to the image irradiance equation

$$I = \omega \frac{1 - Ap - Bq}{\sqrt{1 + p^2 + q^2}}$$

where I is the intensity, (p,q) is the gradient of the surface point whose image has intensity I, ω is the albedo at that point and $(A,B,-1)$ the direction of the light source (Horn, 1977; Ikeuchi and Horn, 1981). We call

$$R(p,q) = \frac{S_w}{d^2} \frac{1 - Ap - Bq}{\sqrt{1 + p^2 + q^2}}$$

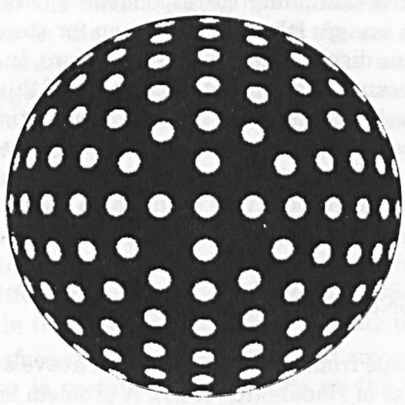


Figure 20. Image of a patterned sphere.

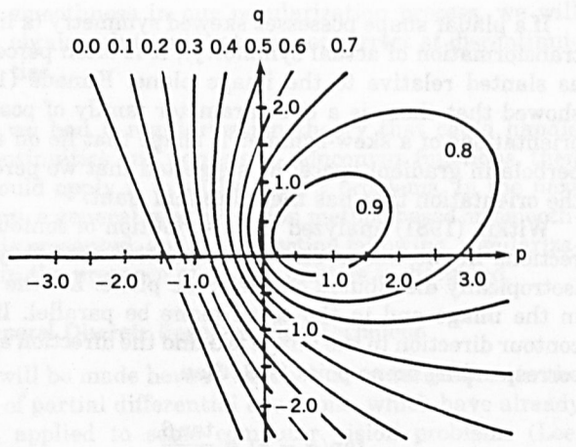


Figure 21. Textural reflectance map.

the "textural reflectance." If we fix S_w/d^2 and the position (A,B) of the texel on the image, this equation can be graphed conveniently as a series of contours of constant textural intensity. Figure 21 illustrates a simple textural reflectance map. Using $R(p,q)$, we can recover shape in a region Ω in the same way as Ikeuchi and Horn recovered shape from shading and occluding boundaries, ie, by minimizing an expression of the form

$$\int\int_{\Omega} \left\{(S_I - R)^2 + \frac{\lambda}{d^2}(p_x^2 + p_y^2 + q_x^2 + q_y^2)\right\} dxdy,$$

with λ a constant weighing the relative importance of the constraint vs. smoothness. Results obtained using this method are shown in Figure 22.

Shape from Contour

"Shape from contour" refers to methods of inferring surface orientation from the shapes or orientations of planar contours (edges or lines) in the image. Perceptually, shape from contour seems to be significantly more powerful than shape from texture (Braunstein and Payne, 1969) or shape from shading (Barrow and Tenenbaum, 1981a).

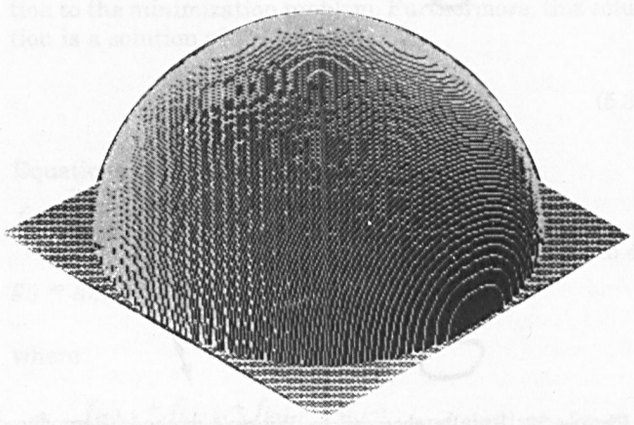


Figure 22. Reconstructed sphere (from Figure 20).

If a planar shape possesses skewed symmetry (a linear transformation of actual symmetry), it is often perceived as slanted relative to the image plane. Kanade (1981) showed that there is a one-parameter family of possible orientations of a skew-symmetric shape that lie on a hyperbola in gradient space; he suggested that we perceive the orientation that has the minimum slant.

Witkin (1981) analyzed the distribution of contour directions in the image on the assumption that they are isotropically distributed on the scene plane. Let the axes in the image and in the scene plane be parallel. If the contour direction in the image is α and the direction at the corresponding scene point is β, then

$$\tan(\alpha - \tau) = \frac{\tan\beta}{\cos\sigma}$$

where σ,τ are the slant and tilt of the scene plane. If measurements are aggregated from the whole image then a distribution of contour directions α can be constructed. One can evaluate the likelihood of this observed distribution of α, given expected distributions for β, σ, and τ. Witkin shows that the probability density function of σ is $\sin\sigma/\pi^2$. If we assume that τ and β are uniformly distributed, it can be shown that the probability of (σ,τ) given the set of measurements α_i is

$$\prod_{1\le i\le n} \frac{\pi^{-2}\sin 2\sigma}{2(\cos^2(\alpha_i - \tau) + \sin^2(\alpha_i - \tau)\cos^2\sigma)}$$

The maximum likelihood estimate for surface orientation is the value (σ,τ) that maximizes this probability. Figure 23 demonstrates the results of this method applied to a variety of shapes and compares it to human estimated tilt. This method can also be applied to the problem of shape from texture under the assumption that contour directions are isotropically distributed.

Brady and Yuille (1984) proposed a general paradigm in which the assumed surface orientation is the one that extremizes some function computed on the scene contour(s). One possible function is $\oint k^2 ds$, where k is the curvature of the contour; minimization of this function has been used as a criterion for interpolating across gaps in plane curves (Horn, 1981). However, this function is not extremized when we transform an ellipse into a circle, whereas ellipses are often perceived as slanted circles. A related function proposed by Barrow and Tenenbaum (1981b) is $\oint (dk/ds)^2\, ds$. However, this function is sensitive to noise, since it involves derivatives. It is also biased toward slants close to 90°.

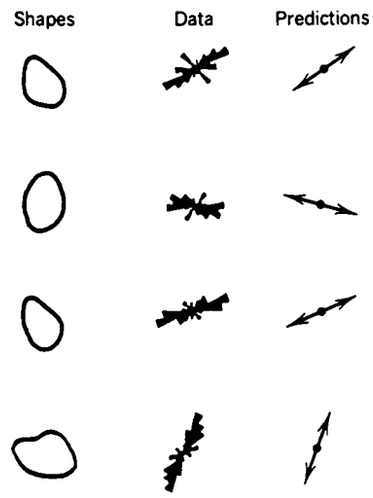


Figure 23. Results obtained by Witkin's method. (Reproduced from Witkin (1981).)

Brady and Yuille used the function

$$m = \frac{\text{area}}{(\text{perimeter})^2}$$

Given an image contour, we choose the orientation for which the scene contour maximizes m. When this is done, an ellipse is interpreted as a slanted circle, a parallelogram as a rotated square, a triangle as a slanted equilateral triangle, and a skewed symmetric figure as symmetric. For other recent research on geometric interpretation of image contours see (Horaud and Brady, 1987).

When there is specific a priori knowledge about the scene contour (perimeter, area, etc), unique solutions for surface orientation can be obtained (Augusteijn and Dyer, 1986; Chou and co-workers, 1987). Another important topic, which has not been treated here, is the analysis of line drawings of 3-D surfaces. The interested reader can consult the papers of Malik (1987), Nalwa (1987) and Koenderink (1986) as well as their references.

Shape from Stereo and Shape (or Structure) from Motion

Given two images of a scene taken by two cameras whose relative position and orientation is known, if corresponding points can be found in the two images (ie points which are the projection of the same scene point), then by the process of triangulation the depth of the scene points can be computed. If many pairs of corresponding points that lie on the same surface can be found, the shape of this surface can be determined. This process is known as stereo (from a Greek word meaning *solid*). If the relative position and orientation of the cameras is not known, then finding the shapes of the surfaces in the scene from corresponding points in the two images is known as the problem of shape (or structure) from motion. In both cases, research has concentrated on finding correspondences between points in the two images (the disparity map for stereo, and the optic flow or displacement field for motion). In the case of motion, because of the great importance of this module in navigation, there has been extensive theoretical development. These topics will not be treated here since they are discussed in detail in separate articles in this encyclopedia which deal respectively with stereo vision and visual motion analysis.

ILL-POSEDNESS AND REGULARIZATION

All the shape from x problems treated above are ill-posed in the sense of Hadamard (1923). A problem is well-posed when its solution exists, is unique, and depends continu-

ously on the given data. Ill-posed problems fail to satisfy one or more of these criteria.

Poggio and his colleagues (Poggio and Koch, 1984; Poggio and Torre, 1984) realized that most recovery problems are ill-posed and that regularization theory (Tichonov and Arsenin, 1977; Morozov, 1984) can be used for their solution. The main approach to "solving" ill-posed problems, ie, restoring "well-posedness," is to introduce suitable constraints that restrict the space of admissible solutions. The problem of finding s (the scene) from i (the image) is ill-posed because the image is obtained from the scene by a noninvertible process, $i = Qs$. Regularization of this problem is usually done by finding the s that minimizes the function $\|Qs - i\|_D^2 + \lambda\|Ps\|_S^2$, where $\| \|_D$ and $\| \|_S$ are norms, P is a functional that usually involves smoothness, and λ is the so-called regularization parameter.

In shape from shading and shape from texture the gradient (p,q) of a surface patch is related to the data at the image point (x,y) which is the projection of that patch by an expression of the form

$$A(x,y)p^2 + B(x,y)q^2 + C(x,y)pq + D(x,y)p + E(x,y)q + F(x,y) = 0 \quad (5.1)$$

where A, B, C, D, E, F are functions of position in the image and depend on the particular physical parameters data). Indeed, in shape from shading the relationship is

$$I(x,y) = \frac{\lambda(p\,p_s + q\,q_s + 1)}{\sqrt{1 + p^2 + q^2}\sqrt{1 + p_s^2 + q_s^2}}$$

where (p_s, q_s) is the direction of the light source, λ the albedo, and $I(x,y)$ the image intensity at (x,y). In shape from texture, similarly, the relationship is

$$S_I = \frac{S_W}{d^2}\frac{1 - Ap - Bq}{\sqrt{1 + p^2 + q^2}}$$

where S_W/d^2 is the "textural albedo," and (A,B) the centroid of the texel. In such shape from x problems the constraint can be rewritten in the form $L(f,g,x,y) = 0\ \forall(x,y)$ on the image plane, where (f,g) is the orientation of the corresponding 3-D point in stereographic coordinates whose space is bounded (Horn, 1986).

In general, suppose the surface orientation satisfies an equation of the form $L(f,g,x,y) = 0$, where x,y are the coordinates in the image plane and f,g are the stereographic coordinates of the normal to the surface patch whose image is at (x,y). Can regularization theory be used to solve for (f,g) everywhere in the image plane, perhaps with the help of boundary conditions? If this is attempted, we face the following problem:

- The equation $L = 0$ is nonlinear, but standard regularization deals with linear constraints. The situation is different if L is convex (Morozov, 1984), but it is not in the cases of shape from shading or texture.
- The surfaces in the scene are not all smooth, and the scene is certainly not smooth at the boundaries where one surface occludes another. If we use smoothness in our regularization process, we will obtain solutions that are not correct at discontinuities.

If we had a regularization theory that could handle discontinuities and nonlinear, nonconvex functions, then we could apply it to shape from x problems. In the next section, a general regularization method based on smoothness is presented, and in the section following, regularization in the presence of discontinuities is discussed.

A General Discrete Regularization Technique

Use will be made here of widely known techniques in the area of partial differential equations, which have already been applied to some computer vision problems (Lee, 1985). Consider the equation $L(f,g,x,y) = 0$, for $(x,y) \in D$, where D is a compact region in the x–y plane. The problem is discretized using an $m \times m$ grid and difference operators instead of differential operators and sums instead of integrals. It is assumed that the surface normals on the boundary of D are known. The desired surface is the one that minimizes an expression of the form

$$e = \sum_{i,j}(s_{ij} + \lambda l_{ij}) \quad (5.2)$$

where the sum is taken over all grid points (i,j) that do not lie on the boundary. Let (f_{ij}, g_{ij}) be the surface orientation at the grid point (i,j). If (i,j) is not a boundary point, $f_{i+1,j}$, $f_{i,j+1}$, $g_{i+1,j}$ and $g_{i,j+1}$ all exist, and we define the smoothness component of e as

$$s_{ij} = m^2\{[f_{i+1,j} - f_{ij}]^2 + [f_{i,j+1} - f_{ij}]^2 + [g_{i+1,j} - g_{ij}]^2 + [g_{i,j+1} - g_{ij}]^2\}$$

It is assumed here for simplicity that D is square; if it has a different shape a different discrete approximation to $f_x^2 + f_y^2 + g_x^2 + g_y^2$ can be used.) Similarly, we define

$$l_{ij} = [L(f_{ij},g_{ij},i,j)]^2$$

The minimization of e is subject to boundary conditions, since f_{ij} and g_{ij} are known if (i,j) is on the boundary.

e is defined on a compact subset K of R^2, and it is continuous in each f_{ij} and g_{ij}. Therefore, there exists a solution to the minimization problem. Furthermore, this solution is a solution of the system

$$\frac{\partial e}{\partial f_{ij}} = \frac{\partial e}{\partial g_{ij}} = 0. \quad (5.3)$$

Equations 5.3 yield

$$f_{ij} = f_{ij}^* - \frac{1}{4}\lambda m^2[L(f_{ij},g_{ij},i,j)]\frac{\partial L}{\partial f}(f_{ij},g_{ij},i,j)$$
$$g_{ij} = g_{ij}^* - \frac{1}{4}\lambda m^2[L(f_{ij},g_{ij},i,j)]\frac{\partial L}{\partial g}(f_{ij},g_{ij},i,j) \quad (5.4)$$

where

$$f_{ij}^* = \frac{f_{i+1,j} + f_{i,j+1} + f_{i-1,j} + f_{i,j-1}}{4} \text{ and similarly for } g_{ij}$$

This can be written compactly as

$$\Phi\xi = -\lambda m^2 \phi(\xi) \tag{5.5}$$

where

$$\xi = [f_{1,1}, \ldots, f_{1,k}, \ldots, f_{k,k}, g_{1,1}, \ldots, g_{kk}]^T$$

$$\phi = [\cdots, \{L(f_{ij},g_{ij},i,j)\}\frac{\partial L(f_{ij},g_{ij},i,j)}{\partial f}, \ldots, \{L(f_{ij},g_{ij},i,j)\}\frac{\partial L(f_{ij},g_{ij},i,j)}{\partial g}\cdots]^T$$

and

$$\Phi = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix} \text{ where } A = \begin{pmatrix} B & -I & & & \\ -I & B & -I & & \\ & \cdots & & & \\ & & \cdots -I & B & I \\ & & & -I & B \end{pmatrix} \in R^{n\times n}$$

$$\text{and } B = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \cdots & & & \\ & & \cdots -1 & 4 & 1 \\ & & & -1 & 4 \end{pmatrix} \in R^{k\times k}.$$

Equation 5.5 is a necessary condition on the solution that minimizes 5.2.

It can be proven that equation 5.5 has a unique solution for appropriately chosen λ (Aloimonos, 1988b). Furthermore, the sequence $\xi^{(\alpha)}$ defined by

$$\xi^{(\alpha+1)} = -\lambda m^2 \Phi^{-1}\phi(\xi^{(\alpha)}), \alpha = 0, 1, 2, \cdots$$

converges to this unique solution. There thus exists a unique surface minimizing e, which is also the unique solution of equation 5.5, and the above described algorithm converges to that solution. A similar result holds even if we do not know (f,g) on the boundary, as long as we assume that "natural" smoothness conditions hold on the boundary.

Discontinuous Regularization

All the shape from x problems come under the regularization paradigm and so regularization is very appealing as a theory for all these modules. However, discontinuities appear in the world (and it is the discontinuities that make it interesting), so a theory of regularization effective in the presence of discontinuities is much needed. Most existing theories of discontinuous regularization explicitly search for boundary points: one segments the image into homogeneous regions and performs ordinary regularization in each region. This can also be done iteratively (Schunck (1984), for example, iteratively combines motion estimation and segmentation). Grimson and Pavlidis (1985) suggest not smoothing over regions where local differences of nearby points are larger than the statistics of the data as a whole would lead one to expect. Lee and Pavlidis (1987) and Lee (1986) use post-validation to find points which are likely to be boundary points. But sometimes we need to regularize over heterogeneous regions in order to take advantage of texture information or in order to attain robustness in the presence of noise by regularizing over a large enough region.

There does not exist a rigorous theory of segmentation. A rigorous discontinuous regularization theory might be the first step in the development of such a theory. A regularization paradigm due to Geman and Geman (1984) uses Bayesian statistical theory to obtaln a non-convex variational measure that must be minimized. The variational condition is based on a probability distribution of a point being a boundary point. The work of Geman and Geman has been extended by Marroquin (1984, 1986) and by Mumford and Shah (1985). Geman's and Marroquin's work deals with a lattice of points whereas Mumford's work deals with a continuous domain. Geman's work deals with functions whose range is finite and small (for example, binary variables), while Marroquin's and Mumford's work deals with functions with a continuous range such as the real numbers. All these papers employ optimization procedures such as simulated annealing that seem to work reasonably well but cannot be guaranteed not to be fooled by multiple local minima. Blake and Zisserman (1987) use "graduated non-convexity" to minimize a series of variational measures which gradually approach the desired measure. Such continuation methods are certain to converge, but they are slow (Allgower and Georg, 1980; Chow and co-workers, 1978).

Standard regularization methods also smooth excessively over regions where the change is steep, but is not large enough to indicate a discontinuity. This suggests that λ, which weighs the importance of smoothing versus consistency with the data, should vary with position. Terzopoulos (1984) has pursued this idea; his smoothness term is a weighted sum of squares of first- and second-order derivatives, where the weights vary with position. However, he primarily investigated the case where the weights are constant except at a small fraction of points where they are zero. Nagel's "oriented smoothness" paradigm (Nagel and Enkelmann, 1986) is another kind of discontinuous regularization.

Shulman and Aloimonos (1988b) developed a method of regularization that does not smooth over discontinuities and does not make a rigid binary distinction between discontinuities and nondiscontinuities. It uses quadratic variational conditions, which yield linear equations.

The basic insight of this method is that we can expect the errors at nearby points to be correlated. Thus $\partial L/\partial x$ and $\partial L/\partial y$ should be small. In addition, a quadratic measure of smoothness such as $\int[(f_x)^2 + (f_y)^2 + (g_x)^2 + (g_y)^2]$ excessively penalizes large changes in orientation, and so produces large jumps in L in the vicinity of discontinuities. These jumps can be controlled by requiring smallness of the derivatives of L, and also by using a more general measure of smoothness.

The general minimization problem that our method solves has the form

$$\text{minimize} \int \left[\sum_0^\infty a_{ij}(\partial^{i+j}L/\partial x^i \partial y^j)\right]^2 dxdy + \int \sum_0^\infty b_{ij}(\partial^{i+j}f/\partial x^i \partial y^j)^2 + \int \sum_0^\infty c_{ij}(\partial^{i+j}g/\partial x^i \partial y^j)^2.$$

where the coefficients are parameters. Note that we impose a requirement of smallness on the derivatives of all orders. Of course, most of the coefficients can be 0. In fact, because of the noisiness of high-order derivative estimates, the coefficients should rapidly approach 0 as $i,j \to \infty$.

A problem with this approach is that computing the derivatives of L requires calculating derivatives of the data, and this calculation is numerically unstable (Poggio and co-workers 1985). As Poggio suggested in connection with the problem of edge detection, in order to differentiate numerically regularization is used. It is unnecessary to actually compute the derivatives; all one needs to know is that they can be approximated by linear functionals of L. Thus, the first integral in our condition is approximated by an expression of the form $(AL)^2$, where A is a matrix. Similarly, the second and third integrals are approximated by a polynomial that is quadratic in the f_{ij} and g_{ij}. We thus obtain Euler-Lagrange equations of the form

$$A(L)\frac{\partial A(L)}{\partial f} = -\Phi_1\xi$$

$$A(L)\frac{\partial A(L)}{\partial g} = -\Phi_2\xi$$

Here $A(L)$ is a sum of the form $\Sigma a_{ij}A^{ij}(L)$ where $A^{ij}(L)$ is an approximation to $\partial^{i+j}L/\partial x^i\partial y^j$. Thus, finally the following is obtained:

$$\sum a_{ij}^2 A^{ij}(L)\partial A^{ij}(L)/\partial f = \Phi_1\xi$$

$$\sum a_{ij}^2 A^{ij}(L)\partial A^{ij}(L)/\partial g = \Phi_2\xi$$

Write $\phi_{ij}(\xi)$ for the known function $\begin{pmatrix} A^{ij}(L)\partial A^{ij}(L)/\partial f \\ A^{ij}(L)\partial A^{ij}(\mathrm{L})/\partial g \end{pmatrix}$; then

$$\sum a_{ij}^2\phi_{ij}(\xi) = \Phi\xi$$

More generally, the a_{ij}^2 could be matrices rather than constants (the constraints that the derivatives of L be small could be relaxed at certain points). If $a_{00} = I$ and $a_{ij} = 0$ for $i,j > 0$ our equations become

$$L\partial L/\partial f = -\Phi_1\xi$$
$$L\partial L/\partial g = -\Phi_2\xi$$

which is the usual (nondiscontinuous) regularization condition. Rewrite $\Sigma a_{ij}^2\phi_{ij}(\xi) = -\Phi\xi$ as

$$\left(\sum \hat{a}_{ij}^2\phi_{ij}(\xi)\right) + \Phi\xi = -\phi_{00}(\xi) \qquad (5.17)$$

where $\hat{a}_{00} = I$ and $\hat{a}_{ij} = a_{ij}$ otherwise. The first term in 5.17 represents the discontinuous correction to the usual regularization condition and we want this term to be as small as possible. To make it easier to work with, rewrite 5.17 in the form

$$\Gamma\Xi = -\phi_{00}(\xi) \qquad (5.18)$$

where Ξ is the vector $[\phi_{00}(\xi),\phi_{01}(\xi),\phi_{10}(\xi), \ldots, \phi_{mn}(\xi)]^T$ and Γ is a matrix. We choose Γ to be the least squares solution of 5.20. Computing this solution involves calculating the Moore-Penrose inverse of Ξ (Penrose, 1955; Ben-Israel and Greville, 1974). Since this is a very complex calculation, it can instead be calculated as the best solution in a restricted subspace. Note that Γ hides the regularization parameter λ. This parameter might need to vary from place to place. (We might require a different amount of smoothing near the boundary than near the center of the visual field.) Our smoothing condition might involve a combination of derivatives of different orders. Γ weights the relative importance of the various derivatives of L being small. This too can vary with position.

If Γ can be found (through adaptive estimation from examples, for instance), the resulting discontinuous regularization technique can be applied to solve various recovery problems. Since the equations involved may be nonlinear, reliable methods of solution still need to be developed. For some preliminary work on recovery tasks using discontinuous regularization see (Aloimonos and Shulman, 1989a; Hurlbert and Poggio, 1987).

MULTIPLE CUES

The methods described in the section on Shape from x are summarized in Figure 24. In each of these methods, shape is computed from a single cue; cues are not combined. Deriving shape from one cue leads to ill-posed problems, with all their associated difficulties. The situation is much improved when we try to recover shape from two or more cues. Indeed, shape can be computed uniquely from many pairs of cues. This is shown in Figure 25. Various examples of this can be found in Aloimonos (1986). Recent research along these lines includes the work of Horn (1986) on combining shading with contour, of various researchers (Richards, 1986; Huang and Blostein, 1985) on combining stereo and motion, of Grimson (1984) on combining shading and stereo, and of Milenkovic and Kanade (1985) on trinocular stereo. In this section a few examples of the use of multiple cues in shape recovery are given.

To analyze shape from shading and retinal motion, assume that the Lambertian surface of an object moves and the optic flow (displacement) field $(\Delta u(x,y), \Delta v(x,y))$ is

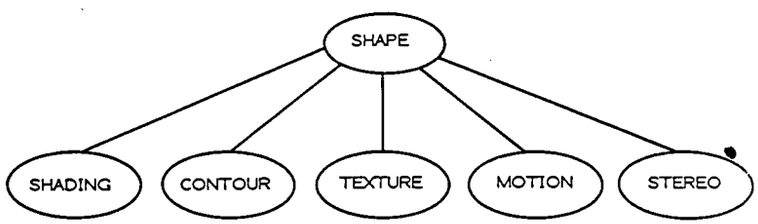


Figure 24. Previous status of shape from x research.

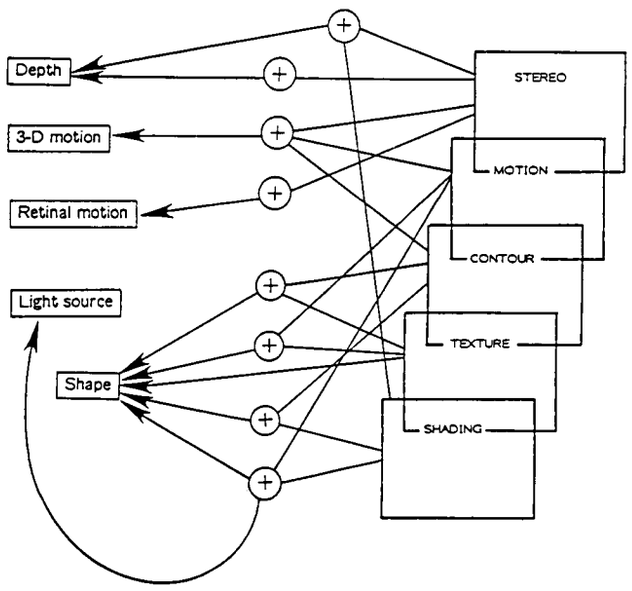


Figure 25. Combining cues.

available everywhere in the image. Then the following constraints hold (Aloimonos and Basu, 1987):

1. The image irradiance equation.
2. A constraint of the form

$$A_1p^2 + B_1q^2 + C_1pq + D_1 = 0,$$

where the coefficients are functions of the displacements.

3. A constraint due to image irradiance, light source and motion, which is of the form

$$A_2p^2 + B_2q^2 + C_2pq + D_2p + E_2q + F_2 = 0$$

Here the coefficients are a function of the displacements, light source position and intensities. Constraints (b) and (c) are constructed on the basis of geometric and photometric invariants. Figure 26 shows a schematic description of the constraints and the solution as the point of intersection of all the constraints.

To analyze shape from contour in multiple images, let C be a contour on the scene plane with equation $Z = pX + qY + c$, and let C_l and C_r be the projections of C on the two images, using perspective projection (Fig. 27).

The difference between the areas or perimeters of C_l and C_r depends on the orientation of the scene plane. If S_L and S_R are the areas of C_l and C_r, it can be shown that

$$\frac{S_L}{S_R} = \frac{1 - A_Lp - B_Lq}{1 - A_Rp - B_Rq}, \tag{6.1}$$

where (A_L, B_L), (A_R, B_R) are the centers of mass of the left and right image contours respectively and the focal length is unity (Aloimonos and Swain, 1988). This gives a linear equation in p,q. If there are more than two images (whose centers are not collinear), we can get additional linear equations and (over)determine (p,q). Alternatively, a constraint involving perimeter can be used. For any contour C_i on the image plane, the corresponding scene plane contour has perimeter

$$\int_C \sqrt{E\,dx^2 + 2F\,dx\,dy + G\,dy^2} \tag{6.2}$$

where E, F, G are the first fundamental coefficients of the mapping from the image to the scene (Lipschutz, 1969). Expression 6.2 can be used to compute the length of the scene contour C from either of its projections C_l and C_r. Since these lengths should be the same, equating their difference to zero yields a nonlinear constraint on the parameters of the scene plane. This, together with equation

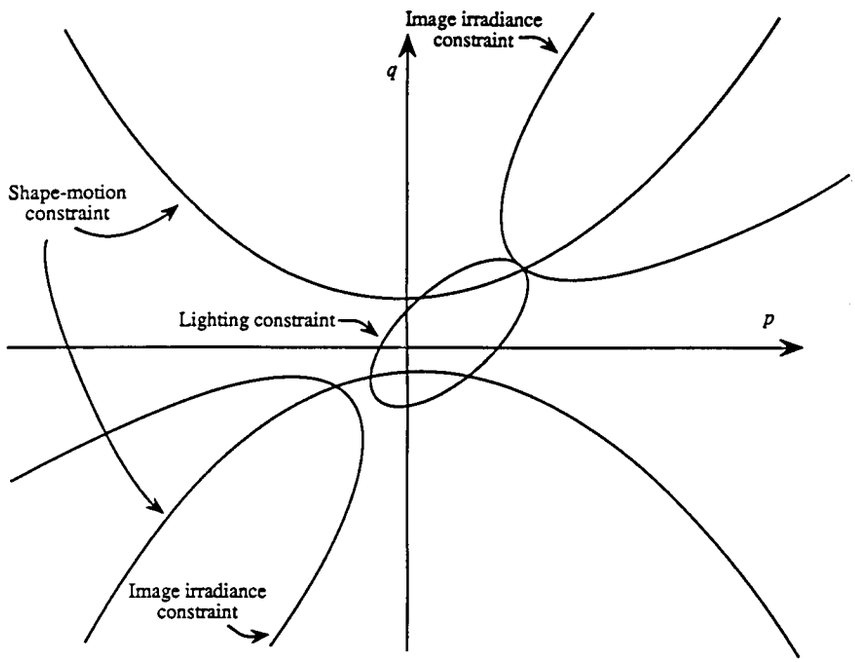


Figure 26. Intersection of constraints.

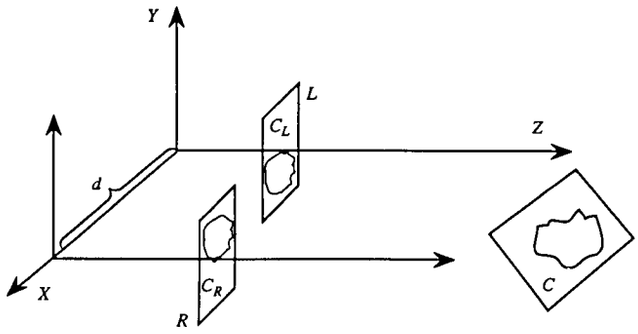


Figure 27. Contours of two images.

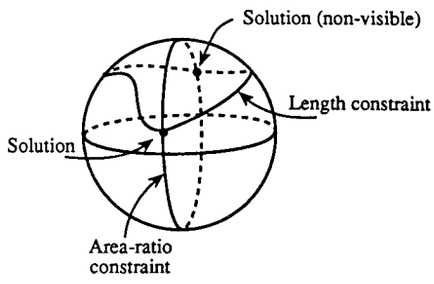


Figure 28. Constraints on the Gaussian sphere.

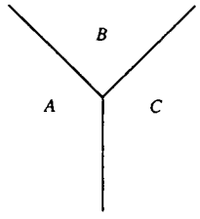


Figure 29. Trihedral vertex.

6.1, gives a finite number of solutions for the orientation (p,q) of the scene plane.

If the equation 6.1 is transformed to the coordinates of the Gaussian sphere, it represents a great circle G. We need to find which point of that great circle satisfies the perimeter constraint. Figure 28 shows the area and perimeter constraints drawn on the Gaussian sphere. It was recently shown (Aloimonos and Hervé) that there can be at most two solutions; and criteria for checking the multiplicity of the solutions have been developed.

Finally, the recovery of shape from shading and contour is illustrated with a simple example from the domain of polyhedra. Consider a trihedral vertex (three planes A, B, C intersecting at a point). Given its orthographic projection (Fig. 29) one would like to recover the orientations of the planes. If there is no other information, the only thing that can be concluded is that the gradients (p_A, q_A), (p_B, q_B) and (p_C, q_C) of A, B, and C form a triangle in gradient space, but the shape of this triangle is unknown (Shafer and co-workers, 1983). If there is also shading information (Horn, 1977), and it is assumed that the planes all have the same albedo, a finite number of solutions for the orientations can be found, using the additional constraints imposed by the image irradiance equation. If it is assumed that the dihedral angles between the planes are known, again only a finite number of solutions is possible. To find these solutions, a triangle must be found in gradient space whose vertices lie on the conic sections that result from the shading or dihedral constraints.

ACTIVE VISION

In this section the active vision paradigm is introduced, in which the observer controls the geometric parameters of the sensor—for example, its position, its orientation, its focal length, etc. This allows the observer to manipulate the constraints on the image(s) and thus provide additional information for solving recovery problems. This paradigm is partly motivated by human and animal perception, which are active. Perceptual activity is exploratory and searching. Humans do not merely see, they actively look (Bajcsy, 1985). When the activity is a known motion of the observer it has been shown (Aloimonos and co-workers, 1988) that all the shape from x problems become well-conditioned and unique solutions become possible. Some of these results are summarized in Table 1.

Table 1. Recovery Problems Are Easier to Solve for an Active Observer.

Problem	Passive Observer	Activer Observer
Shape from shading	Ill-posed problem. Needs to be regularized. Even then, unique solution is not guaranteed because of nonlinearity.	Well-posed and stable. Linear equation; unique solution.
Shape from contour	Ill-posed problem. Has not been regularized up to now in the Tichonov sense. Solvable under restrictive assumptions.	Well-posed problem. Unique solution for either monocular or binocular observer.
Shape from texture	Ill-posed problem. Needs some assumption about the texture.	Well-posed problem. No assumption required.
Structure from motion	Well-posed but unstable. Nonlinear constraints.	Well posed and stable. Quadratic constraints, simple solution methods, stability.

The basis for the active vision approach lies in being able to work in an enriched sensory domain with a partially known parametrization. As the sensor parameters are varied, the image undergoes local transformations that provide powerful constraints for computing the unknown scene parameters. Note that we do not work with a small set of discrete observations, but with continuous trajectories in the stimulus space. These trajectories are smooth, since the sensor transformations are smooth. Thus it is unnecessary to rely on the smoothness of the observed scene. Complications usually associated with multiview approaches to vision—for instance, the correspondence problem—are also avoided.

As an example, consider the active perception of shape from linear image features (texture) (Ito and Aloimonos, 1987). Suppose that a moving camera is looking at a planar surface. To simplify our analysis it is assumed, equivalently, that the surface is moving. Let $\mathbf{X} = (X,Y,Z) \in S$ be imaged onto $\mathbf{x} = (x,y)$ in the image plane R. Let the motion consist of a translation $\mathbf{T} = (T_1,T_2,T_3)$ and a rotation $\Omega = (\omega_1,\omega_2,\omega_3)$, so that $\mathbf{V(X)} = T + \Omega \times \mathbf{X}$, where $\mathbf{V(X)}$ is the velocity of X. Then

$$\mathbf{V(X)} = \sum_{k=1}^{6} r_k V_k(\mathrm{X}), \text{ where} \tag{7.1}$$

$$r_1 = T_1,\ V_1(\mathbf{X}) = (1\ 0\ 0)^T;\ r_4 = \omega_1,\ V_4(\mathbf{X}) = (0\ -Z\ Y)^T$$
$$r_2 = T_2,\ V_1(\mathbf{X}) = (0\ 1\ 0)^T;\ r_5 = \omega_2,\ V_5(\mathbf{X}) = (Z\ 0\ -X)^T$$
$$r_3 = T_3,\ V_3(\mathbf{X}) = (0\ 0\ 1)^T;\ r_6 = \omega_3,\ V_6(\mathbf{X}) = (-Y\ X\ 0)^T$$

It can be easily proved that the image velocity (u,v) at $\mathbf{x} = (x,y)$ is

$$\dot{\mathbf{x}} = \sum_{k=1}^{6} r_k \mathbf{u}_k(\mathbf{x}) = \sum_{k=1}^{6} r_k \begin{bmatrix} u_k(\mathbf{x}) \\ v_k(\mathbf{x}) \end{bmatrix},$$

where $\mathbf{u}_k$ depends on the orientation of S.

For perspective projection the parameters in equation 7.1 are $r_1 = T_1/c$, $r_2 = T_2/c$, $r_3 = T_3/c$, $r_4 = \omega_1$, $r_5 = \omega_2$, $r_6 = \omega_3$, $u_1(\mathbf{x}) = (1 - px - qy, 0)$, $u_2(\mathbf{x}) = (0, 1 - px - qy)$, $u_3(\mathbf{x}) = (-x(1 - px - qy), -y(1 - px - qy))$, $u_4(\mathbf{x}) = (xy, y^2 + 1)$, $u_5(\mathbf{x}) = (-(x^2 + 1), -xy)$ and $u_6(\mathbf{x}) = (y, -x)$, where the motion is translation $T = (T_1,T_2,T_3)$ and rotation $\Omega = (\omega_1,\omega_2,\omega_3)$ and the scene plane has equation $Z = pX + qY + c$ with respect to the camera coordinate system (Fig. 30).

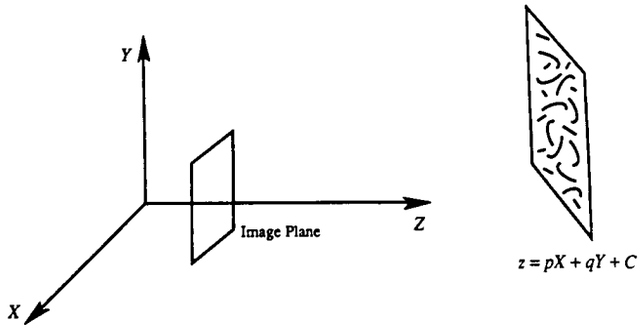


Figure 30. Imaging system and a textured surface.

Let the image intensity be $I(x,y)$. A *linear feature* (Amari, 1987) is a linear functional,

$$f = \int\int I(x,y)m(x,y)dx\,dy,$$

where m is called a measuring function.

The image velocity satisfies the following equation (Horn, 1986):

$$I_x u + I_y v + I_t = 0,$$

where (u,v) is the optic flow at a point (x,y) and I_x, I_y, I_t are the spatiotemporal derivatives of the image intensity function at the point (x,y). This equation can be written as

$$\frac{\partial I}{\partial t} = -\dot{\mathbf{x}} \cdot \nabla I.$$

The time derivative of a linear feature will be

$$\dot{f} = \int\int \frac{\partial I}{\partial t} m\,dx\,dy = -\int\int m(\dot{\mathbf{x}} \cdot \nabla I)dx\,dy$$

$$= -\sum_{k=1}^{6} r_k \int\int m(u_k I_x + v_k I_y)dx\,dy$$

$$= \sum_{k=1}^{6} r_k h_k, \text{ with } h_k = -\int\int m(u_k I_x + v_k I_y)dx\,dy$$

This equation relates linear features to shape and motion parameters. Furthermore, it is linear. If it is applied to a set of linear features, a set of linear equations in those parameters is obtained. So, a simple linear least-squares method is sufficient for the recovery of the parameters. No local correspondence has been used. The only computed quantities were the time derivatives of linear features, which involve the whole image.

It is also important to note that in this approach, the spatial derivatives of the intensity function do not need to be computed. Integration by parts avoids differentiation of the intensity function; only the derivative of the measuring function has to be computed. This avoids differentiating the image intensity, which is discrete; numerical differentiation is an ill-posed problem. More importantly, the same approach can be followed if the image is discontinuous—for example, a dot pattern or a line pattern (eg edges).

Now the algorithm can be summarized for the active detection of shape from two images $I(x,y,t)$ and $I(x,y,t + dt)$ taken by a camera with known motion.

1. Choose a set of differentiable measuring functions $\mu_i(x,y)$, $i = 1, \ldots, n$. Examples might be $x^i y^j$, $0 \le i,j \le k$, or Fourier features such as $\cos(ix + jy)$, $0 \le i,j \le k$.
2. Compute the linear features $f_i = \int\int I(x,y)\mu_i(x,y)dx\,dy$, where the integration is over any desired area of interest.
3. Estimate the time derivatives of the fs from the images $I(x,y,t)$ and $I(x,y,t + dt)$.

4. Compute $h_k = -\int\int m_i(u_k I_x + v_k I_y)dxdy$ for each f.
5. Let $\mathbf{f}$ be the vector of feature values and H the matrix of hs. From equation $\mathbf{f} = Hr$ solve for p,q (and/or) c using a least squares method.

This method has been used (Aloimonos, 1989) to successfully recover the orientation of planar surfaces containing complex patterns, viewed by a moving camera.

This method involves solving a linear 3×3 system in the unknowns p,q, and c (the parameters of the plane). Let the system be $A\,\tilde{x} = \tilde{c}$, where $\tilde{x} = (p\ q\ c)^T$, $A = (a_{ij})$ is a 3×3 matrix, and $\tilde{c}$ is a 1×3 vector whose components are expressions involving spatiotemporal derivatives. It is possible that the system may be unstable. Since there is a discretization error as well as a slight error in the estimation of the known motion, there is some uncertainty in the elements of the matrix A and the vector c. Let the true system be $A^*\tilde{x}^* = c^*$, and let

$$a_{ij}^* \in [a_{ij} - \varepsilon_{ij}, a_{ij} + \varepsilon_{ij}]$$

If there exist values of the coefficients a_{ij} in these intervals of uncertainty for which the determinant of the system becomes zero, then the system is very badly conditioned and its solution will be unreliable. It can be shown (Kuperman, 1971) that the necessary and sufficient condition for the system not to be critically ill-conditioned is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} |b_{ji}|\varepsilon_{ij} < 1$$

where $(b_{ji}) = A^{-1}$. This expression can be used to test the robustness of the algorithm. Note that this represents a worst case analysis.

RESEARCH GOALS

Research on shape from x has been extensively pursued and has accomplished a great deal. The study of modules that may correspond to specific abilities of the human visual system, along with the formulation and exploitation of photometric and geometric relations, has contributed to the foundations of vision as a scientific field. Rigorous theories have been developed for deriving various intrinsic scene properties from various image characteristics. Most of these theories have found no practical applications because the theories do not result in robust algorithms. Any proposed theory explaining visual abilities must be backed up with a thorough theoretical stability analysis. A theory cannot be used for practical applications or to explain human visual capabilities if it is not robust. The algorithmic level of the Marr paradigm must be accompanied with careful theoretical error analyses. Of course, such analysis is hard. This issue is a topic of research that should be pursued; a few researchers have recently made such attempts (Aloimonos, 1986; Horn and Weldon, 1987; Adiv, 1985; Huang and Blostein, 1987).

Techniques have been developed in numerical analysis (Kuperman, 1971; Neumaier, 1984; Gay, 1981, 1982; Moore and Kioustelidis, 1980; Demmel, 1987) that may be used to study the sensitivity of vision algorithms. One could go further and do a probabilistic analysis, given assumptions about the probability distributions of the input measurements. The assumptions could be tested using statistical techniques.

Most of the shape from x modules have been studied in isolation: (shape from x and/or y). Unification of existing approaches to a given module is also of interest; some research has been done in this direction (Moerdler and Kender, 1987). But a formal theory is needed of how to combine information from different sources, especially, contradictory information. Discontinuous regularization and Markov random field methods are useful tools for this purpose (Marroquin and co-workers, 1985; Gamble and Poggio, 1987; Poggio and co-workers, 1987; Aloimonos and Shulman, 1989b).

Work on *active* vision will lead to further research on *exploratory* and *feedback* vision. Exploratory vision involves determining the activity that yields the most stable algorithm for the task at hand. Feedback vision deals with how information gathered from the environment can be used to guide future activities.

NEW DIRECTIONS

This article has dealt primarily with computer vision as a general recovery problem. Over the past 15 years many elegant mathematical theories describing various recovery modules have been formulated. Unfortunately, very few vision systems perform well in real-world environments. There seem to be several reasons for this.

One reason is that extracting useful visual information from images seems to involve a very large amount of computation. The visual cortexes of animals contain millions of neurons, which perform computations that require very large numbers of computer operations to simulate.

A second reason is the belief (Nelson, 1988) that practical results will eventually flow from a successful theory rather than vice versa. This may have more to do with the scarcity of practical systems than with philosophical conviction; historically, empirical engineering applications or unexplained observations have preceded theoretical developments at least as frequently as the reverse. If machine vision systems suddenly appeared that operated robustly in real-world domains, it is quite likely that theories explaining their commonality would soon follow.

A third reason is that the generally accepted goals for vision systems may be misplaced, or at least over-ambitious. The two commonly held touchstones for practical vision systems, recognition and navigation, are high-level objectives. If both were achieved, computer vision systems would have many of the capabilities of the human visual system. Given the lack of success in developing general systems that realize either of these goals in a robust manner, it would appear reasonable to consider simpler problems. Many researchers have gone in this direction by working on specific industrial applications. However, this work does not enhance our understanding of vision in general.

A more fruitful approach is to address specific classes of vision tasks. For example, the shape from x theories can be applied to obstacle avoidance, but we can also work on obstacle avoidance as a problem in its own right, develop a computational theory for it, design algorithms, prove that they are robust, and implement them. Nelson and Aloimonos (1987) is an example of this approach. Aloimonos and Shulman (1989b) describe this approach as working bottom-up in the Marr paradigm to find general solutions to specific problems.

If we could recover the scene we would be able to perform many tasks, but it is not always necessary to do complete recovery. What is vision for? (Ballard, 1989). Vision is needed in order to accomplish tasks that are essential for our survival: recognize mates, friends, enemies, and food, avoid danger, and so on. To carry out these tasks, it may not be necessary to recover the entire scene and all its properties. When visual abilities are studied, their purposes and their uses should be kept in mind. When vision is studied from a purposive, utilitarian (Ramachandran, 1989), or animate (Ballard, 1989) viewpoint, the problems that formulated are generally simpler, since they are relevant to specific tasks, and as a result they can often be solved by qualitative, robust techniques (Aloimonos, 1990; Zucker, 1988).

Many neuroscientists believe that the visual capabilities of animals developed (evolved) because of specific needs. Some of these abilities were based on common principles, but they may have developed at different times and may be implemented in separate hardware. The parts of the brain devoted to vision seem to implement independent processes (which of course communicate) that are devoted to the solution of specific visual tasks.

Research on recovery can continue, to try to understand why existing approaches are unstable, and try to develop provably optimal methods. A more radical idea would be to reconsider the need for the recovery paradigm. Rather, we should ask what tasks could be performed if there were adequate recovery modules. After these tasks have been identified, we can try to solve them directly, and not treat them as applications of a general recovery process. For example, to avoid obstacles it is necessary to

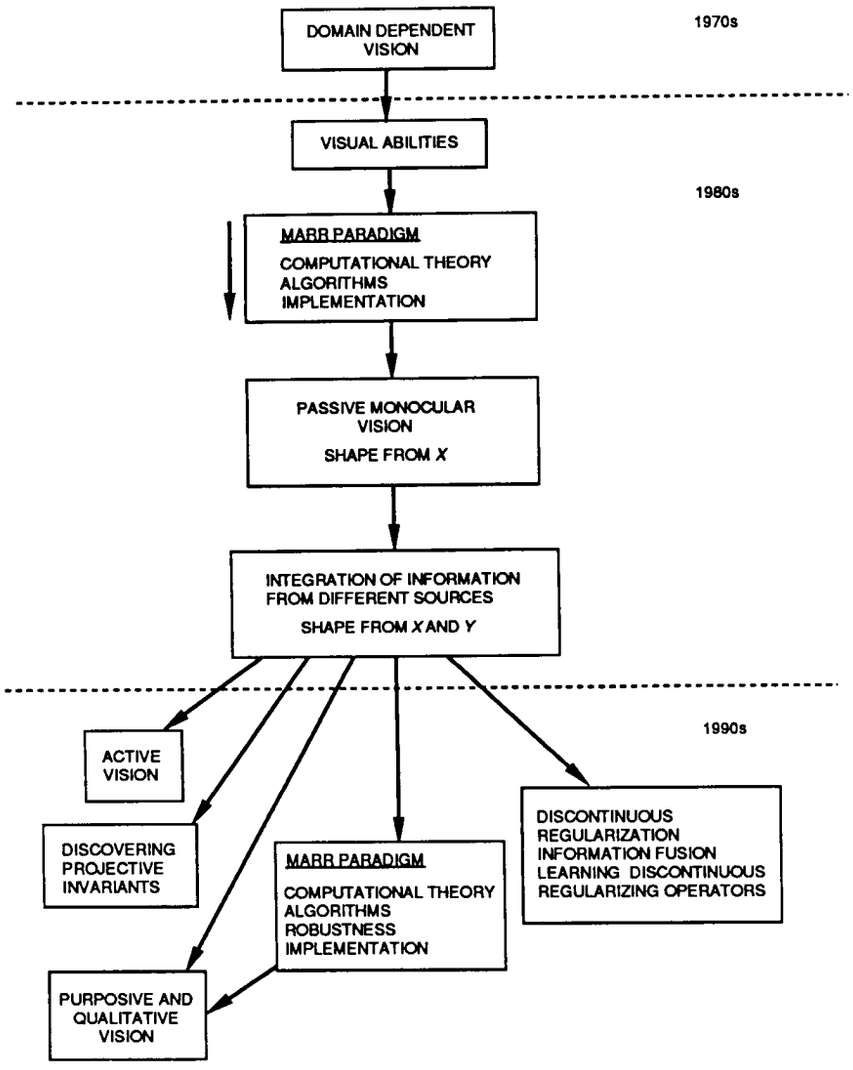


Figure 31. Research trends in visual recovery.

Table 2. Reconstructionist vs Purposive Vision.

Reconstructionist Vision	Purposive Vision
Reconstruct properties of the scene from images.	Define vision-based tasks.
Develop methods of recovering specific properties.	Develop methods of decomposing tasks into simpler subtasks.
Quantitative methods.	Qualitative methods.
General-purpose: recover the scene.	Directed (purposive).

answer a set of specific questions: Is another object present? Is it coming closer to me? If so, will it hit me? If so, how long will it take (relative to my reaction time)? If problems can be solved directly, the structure from motion module may no longer be needed. Moreover, because simple, qualitative questions that have small numbers of possible answers are being asked, it may be possible to achieve robust solutions.

In this framework, one need no longer regard vision as a collection of modules whose purpose is to reconstruct the scene and its properties and thus provide information for accomplishing various tasks. Rather, it can be regarded as a collection of processes which (individually or in groups) solve particular visual tasks. This means that vision is being considered not in isolation (as the recovery school of thought does), but as a part of a system that performs various tasks.

If we wish to study vision in general, we should study the tasks that organisms possessing vision can accomplish. If these tasks are complicated, we should decompose them into simpler tasks, and solve the simpler ones. This will then solve all the subtasks and yield a set of processes which if appropriately combined can perform the original task. This viewpoint is summarized, and contrasted with the reconstructionist viewpoint, in Table 2 (Aloimonos, 1990).

We conclude with Figure 31, which describes the evolution of research in visual recovery and our view about future directions.

BIBLIOGRAPHY

G. Adiv, "Inherent Ambiguities in Recovering 3-D Motion and Structure from a Noisy Flow Field," *Proceedings of the Conference on Computer Vision and Pattern Recognition,* 1985, pp. 70–77.

E. Allgower and K. Georg, "Simplicial and Continuation Methods of Approximating Fixed Points and Solutions to Systems of Equations," *SIAM Rev.* **22,** 28–85 (1980).

J. Aloimonos, Ph.D. thesis, Department of Computer Science, University of Rochester, 1986.

J. Aloimonos, "Shape from Texture," *Biol. Cybernetics* **58,** 345–360 (1988a).

J. Aloimonos, "Visual Shape Computation," *Proc. IEEE,* 899–916 (1988b).

J. Aloimonos, "Unifying Shading and Texture Through an Active Observer," *Proc. R. Soc. London B* **238,** 25–27 (1989).

J. Aloimonos, "Purposive and Qualitative Active Vision," *Proceedings of the DARPA Image Understanding Workshop,* 1990a, pp. 816–828.

J. Aloimonos, "Perspective Approximations," *Image and Vision Computing* **8,** 177–192 (1990b).

J. Aloimonos and A. Basu, "Combining Information in Low-level Vision," *CAR-TR-336,* Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, 1987.

J. Aloimonos and J. Y. Hervé, "Correspondenceless Detection of Depth and Motion for a Planar Surface," *IEEE Trans. PAMI,* in press.

J. Aloimonos and D. Shulman, "Learning Early Vision Computations," *J. Opt. Soc. Am. A* **6,** 908–919 (1989a).

J. Aloimonos and D. Shulman, *Integration of Visual Modules: An Extension of the Marr Paradigm,* Academic Press, Boston, 1989b.

J. Aloimonos and M. Swain, "Shape from Patterns: Regularization," *Intl. J. Comput. Vision* **2,** 171–187 (1988).

J. Aloimonos, I. Weiss, and A. Bandopadhay, "Active Vision," *Intl. J. Comput. Vision* **2,** 333–356 (1988).

S. Amari, Personal communication, 1987.

D. Arnold, "Local Context in Matching Edges for Stereo Vision," *Proceedings of the DARPA Image Understanding Workshop,* 1978, pp. 65–72.

J. F. Augusteijn and C. R. Dyer, "Recognition and Recovery of the Three Dimensional Orientation of Planar Point Patterns," *Comput. Vision, Gr. Im. Process.* **36,** 76–99 (1986).

R. Bajcsy, "Active Perception vs. Passive Perception," *Proceedings of the Workshop on Computer Vision,* 1985, pp. 55–59.

R. Bajcsy and L. Lieberman, "Texture Gradient as a Depth Cue," *Comput. Vision, Gr., Im. Process.* **5,** 52–67 (1976).

R. Bajcsy and F. Solina, "Three Dimensional Object Representation Revisited," *Proceedings of the International Conference on Computer Vision,* 1987, pp. 231–240.

D. H. Ballard, "Reference Frames for Animate Vision," *Proceedings of the International Joint Conference on Artificial Intelligence,* Morgan Kaufmann, San Mateo, Calif., 1989, pp. 1635–1641.

H. G. Barrow and J. M. Tenenbaum, "Computational Vision," *Proc. IEEE* **69,** 572–595 (1981a).

H. G. Barrow and J. M. Tenenbaum, "Interpreting Line Drawings as Three Dimensional Surfaces," *Artif. Intell.* **17,** 75–116 (1981b).

A. Ben-Israel and T. Greville, *Generalized Inverses, Theory and Applications,* John Wiley and Sons, Inc., New York, 1974.

A. Blake, "Specular Stereo," *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* Los Angeles, Morgan-Kaufmann, San Mateo, Calif., 1985, pp. 973–976.

A. Blake and S. Zisserman, *Visual Reconstruction,* MIT Press, Cambridge, Mass., 1987.

M. Brady, "Computational Approaches to Image Understanding," *ACM Comput. Surv.* **14,** 3–71 (1982).

M. Brady and A. Yuille, "An Extremum Principle for Shape from Contour," *IEEE Trans. PAMI* **PAMI-6,** 288–301 (1984).

M. L. Braunstein and J. W. Payne, "Perspective and Form Ratio as Determinants of Relative Slant Judgments," *J. Exper. Psych.* **3,** 584–590 (1969).

M. J. Brooks, "Surface Normals from Closed Paths," *Proceedings of the Sixth International Joint Conference on Artificial Intelligence,* Tokyo, Morgan-Kaufmann, San Mateo, Calif., 1979, pp. 98–101.

M. J. Brooks and B. K. P. Horn, "Shape and Source from Shad-

ing," *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* Los Angeles, Morgan-Kaufmann, San Mateo, Calif., 1985, pp. 932–936.

A. R. Bruss, "The Image Irradiance Equation: Its Solution and Application," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Mass., 1980.

C. T. Chou, J. Aloimonos, and A. Rosenfeld, "Correspondenceless Model Based and Active Perception of Shape from Contour," *CAR-TR-275,* Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, 1987.

S. N. Chow, J. Mallet-Paret, and J. A. Yorke, "Finding Zeros of Maps: Homotopy Methods that are Constructive with Probability One," *Math. Comp.* **32,** 887–899 (1978).

E. Coleman and R. Jain, "Obtaining 3D Shape of Textured and Specular Surfaces Using Four-source Photometry," *Comput. Gr., Im. Process.* **18,** 309–328 (1982).

J. Demmel, "The Geometry of Ill-conditioning," *J. Complexity* **3,** 201–299 (1987).

J. A. Feldman, "Four Frames Suffice: A Provisional Model of Vision and Space," *Behav. Brain Sci.* **8,** 265–313 (1985).

R. Frankot and R. Chellappa, "A Method for Enforcing Integrability in Shape from Shading Algorithms," *Proceedings of the International Conference on Computer Vision,* 1987, pp. 118–127.

E. Gamble and T. Poggio, "Visual Integration and Detection of Discontinuities: The Key Role of Intensity Edges," *AI Memo 970,* Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1987.

D. Gay, "Perturbation Bounds for Nonlinear Equations," *SIAM J. Numerical Analysis* **18,** 654–663 (1981).

D. Gay, "Solving Interval Linear Equations," *SIAM J. Numerical Analysis* **19,** 858–870 (1982).

S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. PAMI* **PAM1–6,** 721–741 (1984).

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

E. Grimson, "Binocular Shading and Visual Surface Reconstruction," *Comput. Vision, Gr., Im. Process.* **28,** 19–43 (1984).

W. E. L. Grimson and T. Pavlidis, "Discontinuity Detection for Visual Surface Reconstruction," *Comput. Vision, Gr., Im. Process.* **30,** 316–330 (1985).

J. Hadamard, *Lectures on the Cauchy Problem in Linear Partial Differential Equations,* Yale, New Haven, Conn., 1923.

G. Healey and T. Binford, "Local Shape from Specularity," *Proceedings of the International Conference on Computer Vision,* 1987, pp. 151–160.

R. Horaud and M. Brady, "On the Geometric Interpretation of Image Contours," *Proceedings of the International Conference on Computer Vision,* 1987, pp. 374–382.

B. K. P. Horn, "Obtaining Shape from Shading Information," in *The Psychology of Computer Vision,* P. H. Winston, ed., McGraw-Hill, New York, 1975, pp. 115–155.

B. K. P. Horn, "Understanding Image Intensities," *Artif. Intell.* **8,** 201–231 (1977).

B. K. P. Horn, "The Curve of Least Energy," *AI Memo 610,* Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1981.

B. K. P. Horn, *Robot Vision,* McGraw-Hill, New York, 1986.

B. K. P. Horn and M. J. Brooks, "The Variational Approach to Shape from Shading," *Comput. Vision, Gr., Im. Process.* **33,** 174–208 (1986).

B. K. P. Horn and M. J. Brooks, *Shape from Shading,* MIT Press, Cambridge, Mass., 1989.

B. K. P. Horn and R. W. Sjoberg, "Calculating the Reflectance Map," *Applied Optics* **18,** 1770–1779 (1979).

B. K. P. Horn and E. J. Weldon, "Computationally Efficient Methods of Recovering Translational Motion," *Proceedings of the International Conference on Computer Vision,* 1987, pp. 2–11.

T. S. Huang and M. Blostein, "Robust Algorithms for Motion Estimation Based on Two Sequential Stereo Image Pairs," *Proceedings of the Conference on Computer Vision and Pattern Recognition,* 1985, pp. 518–523.

T. S. Huang and T. Blostein, "Quantization Errors in Stereo Triangulation," *Proceedings of the IEEE International Conference on Computer Vision,* 1987, pp. 325–334.

A. Hurlbert and T. Poggio, "Learning a Color Algorithm from Examples," *AI Memo 909,* Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1987.

K. Ikeuchi, "Shape from Regular Patterns," *Artif. Intell.* **22,** 49–75 (1984).

K. Ikeuchi and B. K. P. Horn, "Numerical Shape from Shading and Occluding Boundaries," *Artif. Intell.* **17,** 141–184 (1981).

E. Ito and J. Aloimonos, "Determining Three Dimensional Transformation Parameters from Images: Theory," *CAR-TR-318,* Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, 1987.

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

T. Kanade, "Determining the Shape of an Object from a Single View," *Artif. Intell.* **17,** 409–460 (1981).

K. I. Kanatani, "Detection of Surface Orientation and Motion from Texture by a Stereological Technique," *Artif. Intell.* **23,** 213–237 (1984).

K. Kanatani and T. C. Chou, "Shape from Texture: General Principle," *Artif. Intell.* **38,** 1–48 (1989).

J. R. Kender, "Shape from Texture: An Aggregation Transform that Maps a Class of Textures into Surface Orientation," *Proceedings of the Sixth International Joint Conference on Artificial Intelligence,* Tokyo, Morgan-Kaufmann, San Mateo, Calif., 1979, pp. 475–480.

J. Kender, Ph.D. thesis, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pa., 1980.

J. J. Koenderink, "An Internal Representation for Solid Shape Based on the Topological Properties of the Apparent Contour," in W. Richards and S. Ullman, eds., *Image Understanding 1986,* Ablex, Norwood, N.J., 1986, pp. 257–285.

J. J. Koenderink and A. J. van Doorn, "Photometric Invariants Related to Solid Shape," *Acta Optica* **27,** 981–996 (1980).

I. Kuperman, *Approximate Algebraic Linear Equations,* Van Nostrand, London, 1971.

E. H. Land and J. J. McCann, "Lightness and Retinex Theory," *J. Opt. Soc. Am.* **61,** 1–11 (1971).

C. H. Lee and A. Rosenfeld, "Improved Methods of Estimating Shape from Shading Using the Light Source Coordinate System," *Artif. Intell.* **26,** 125–143 (1985).

D. Lee, "A Provably Convergent Algorithm for Shape from Shading," *Proceedings of the DARPA Image Understanding Workshop,* 1985, pp. 489–496.

D. Lee, Ph.D. thesis, Department of Computer Science, Columbia University, New York, 1986.

D. Lee and T. Pavlidis, "One-dimensional Regularization with Discontinuities," *Proceedings of the International Conference on Computer Vision,* 1987, pp. 572–577.

M. Lipschutz, "Differential Geometry," *Schaum's Outline Series,* McGraw Hill, New York, 1969.

J. Malik, "Interpreting Line Drawings of Curved Objects," *Int. J. Comput. Vision* **1,** 73–103 (1987).

D. Marr, *Vision,* W. H. Freeman, San Francisco, 1982.

J. L. Marroquin, "Surface Reconstruction Preserving Discontinuities," *AI Memo 792,* Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1984.

J. Marroquin, Ph.D. thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1986.

J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic Solution of Ill-Posed Problems in Computational Vision," *Proceedings of the DARPA Image Understanding Workshop,* 1985, pp. 293–309.

V. Milenkovic and T. Kanade, "Trinocular Vision Using Photometric and Edge Orientation Constraints," *Proceedings of the DARPA Image Understanding Workshop,* 1985, pp. 163–175.

M. Moerdler and J. Kender, "An Integrated System that Unifies Multiple Shape from Texture Algorithms," *Proceedings of the DARPA Image Understanding Workshop,* 1987, pp. 574–580.

R. Moore and T. Kioustelidis, "A Simple Test for Accuracy of Approximate Solutions to (Non)linear Systems," *SIAM J. Numerical Analysis* **17,** 521–529 (1980).

V. A. Morozov, *Methods for Solving Incorrectly Posed Problems,* Springer, Berlin, 1984.

D. Mumford, "The Problem of Robust Shape Descriptions," *Proceedings of the International Conference on Computer Vision,* 1987, pp. 602–606.

D. Mumford and M. Shah, "Boundary Detection by Minimizing Functionals," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1985, pp. 22–25.

H. H. Nagel and W. Enkelmann, "An Investigation of Smoothness Constraints for the Estimation of Displacement Vector Fields from Image Sequences," *IEEE Trans. PAMI* **PAMI–8,** 565–593 (1986).

V. Nalwa, Ph.D. thesis, Department of Computer Science, Stanford University, 1987.

R. Nelson, Ph.D. thesis, Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, 1988.

R. Nelson and J. Aloimonos, "Using Flow Field Divergence for Obstacle Avoidance in Visual Navigation," *CAR-TR-332,* Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, 1987.

A. Neumaier, "New Techniques for the Analysis of Linear Interval Equations," *Linear Algebra and Its Applications* **58,** 273–325 (1984).

Y. Ohta and T. Kanade, "Stereo by Intra- and Inter-scanline Search Using Dynamic Programming," *IEEE Trans. PAMI* **PAMI–7,** 139–154 (1985).

Y. Ohta, K. Maenobu, and T. Sakai, "Obtaining Surface Orientation from Texels Under Perspective Projection," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence,* Vancouver, B.C., Morgan-Kaufmann, San Mateo, Calif., 1981, pp. 746–751.

T. Pavlidis, ed. "Special Memorial Issue for Professor King-Sun Fu," *IEEE Trans. PAMI* **PAMI–8,** 289–404 (1986).

R. Penrose, "A Generalized Inverse for Matrices," *Proc. Cambridge Philos. Soc.* **51,** 406–413 (1955).

A. P. Pentland, "Finding the Illuminant Direction," *J. Opt. Soc. Am.* **72,** 448–455 (1982).

A. P. Pentland, "Local Shading Analysis," *IEEE Trans. PAMI* **PAMI–6,** 170–187 (1984).

A. P. Pentland, "Perceptual Organization and the Representation of Natural Form," *Artif. Intell.* **28,** 293–331 (1986).

T. Poggio and C. Koch, "An Analog Model of Computation for the Ill-posed Problems of Early Vision," *AI Memo 783,* Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1984.

T. Poggio and V. Torre, "Ill-posed Problems and Regularization Analysis in Early Vision," *Proceedings of the DARPA Image Understanding Workshop,* 1984, pp. 257–263.

T. Poggio and co-workers, "MIT Progress in Understanding Images," *Proceedings of the DARPA Image Understanding Workshop,* 1985, pp. 25–39.

T. Poggio and co-workers, "MIT Progress in Understanding Images," *Proceedings of the DARPA Image Understanding Workshop,* 1987, pp. 41–54.

V. S. Ramachandran, *Proceedings of the Workshop on Visual Motion,* 1989.

W. Richards, "Structure from Stereo and Motion," *J. Opt. Soc. Am. A* **2,** 343–349 (1986).

B. Schunck, "Motion Segmentation and Estimation," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Mass., 1984.

S. Shafer, T. Kanade, and J. Kender, "Gradient Space Under Orthography and Perspective," *Comput. Vision, Gr., Im. Process.* **24,** 182–199 (1983).

D. Shulman and J. Aloimonos, "Boundary Preserving Regularization: Theory, Part I," *CAR-TR-340,* Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, 1988b.

K. A. Stevens, "Surface Perception from Local Analysis of Texture and Contour," *AI Memo 512,* Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1980.

K. A. Stevens, "The Visual Interpretation of Surface Contours," *Artif. Intell.* **17,** 47–73 (1981).

T. M. Strat, "A Numerical Method for Shape from Shading from a Single Image," M.S. thesis, Massachusetts Institute of Technology, Cambrige, Mass., 1979.

D. Terzopoulos, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Mass., 1984.

A. N. Tichonov and V. Y. Arsenin, *Solution of Ill-Posed Problems,* Winston and Wiley, Washington D.C., 1977.

A. P. Witkin, "Recovering Surface Shape and Orientation from Texture," *Artif. Intell.* **17,** 17–45 (1981).

L. B. Wolff, "Physical Stereo for Combined Specular and Diffuse Reflection," *Technical Report,* Department of Computer Science, Columbia University, New York, 1986.

L. B. Wolff, "Spectral and Polarization Stereo Methods Using a Single Light Source," *Proceedings of the DARPA Image Understanding Workshop,* 1987, pp. 810–820.

L. B. Wolff, "Surface Curvature and Contour from Photometric Stereo," *Proceedings of the DARPA Image Understanding Workshop,* 1987, pp. 821–824.

R. J. Woodham, "Analyzing Images of Curved Surfaces," *Artif. Intell.* **17,** 117–141 (1981).

S. Zucker, "The Emerging Paradigm of Computational Vision," *Ann. Rev. Comput. Sci.* **2,** 69–89 (1988).

YIANNIS ALOIMONOS
AZRIEL ROSENFELD
University of Maryland

ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE

This extensively revised and expanded *Second Edition* of the *Encyclopedia of Artificial Intelligence* defines the discipline by bringing together the core of knowledge from all fields encompassed by AI. It covers the latest developments in current AI topics such as neural networks, fuzzy logic, machine vision, natural language generation, and many more. Includes:

- Over 450 articles—all entries written expressly for the *Encyclopedia*
- Over 5,000 literature references; 454 illustrations and color photographs
- Over 50% new and revised material
- Exemplary indexing and cross-referencing for easy, complete information access to all topics

Praise for the *First Edition*...

"The *Encyclopedia* is a wonder of clarity and scope: surprisingly easy to read...the clarity is an especially pleasant surprise, considering the articles were all written by AI experts...It's a treasure house of easily accessible knowledge."
—*Language Technology*

"Excellent bibliographies are attached to most of the articles, and diagrams and sketches are clear and helpful. The indexing and cross-indexing are exemplary. As the editor points out, the reader will be led by the extensive cross-references to almost every other article..." —*Artificial Intelligence Reporter*

"The *Encyclopedia* is a first-class piece of work that will be an indispensable part of any AI library." —*Computing Reviews*

"...A tour de force...a truly fantastic encyclopedia which no one in the field of artificial intelligence can afford to be without."
—*Systems Research & Information*

WILEY-INTERSCIENCE
John Wiley & Sons, Inc.
Professional, Reference and Trade Group
605 Third Avenue, New York, N.Y. 10158-0012
New York • Chichester • Brisbane • Toronto • Singapore
ISBN 0 471-50307-X (Two-Volume Set)
ISBN 0 471-50305-3 (Vol. 1)
ISBN 0 471-50306-1 (Vol. 2)

ISBN 0-471-50306-1

9 780471 503064 90000