Motion 7

7.1 MOTION UNDERSTANDING

Motion imagery presents many interesting challenges to computer vision, but static scene analysis received more attention in the 1960's and 1970's. In part, this may have been due to a technical problem: With most types of input media and domains, motion vision input is much more voluminous than static vision input. However, we believe that a more basic problem has been the assumption that motion vision could best be understood (or implemented) as many static frames analyzed very quickly, with results linked up in temporal sequence. This characterization of motion vision is extreme but perhaps illuminating. First, it assumes that vision involves processing static scenes. Second, it acknowledges that massive amounts of data may be required. Third, in it motion understanding degenerates to a postprocessing step which is mostly a matching operation—the differences or similarities between (understood) frames are analyzed and recorded. The extreme "static is basic" view is that motion is an unnaturally complex or difficult problem because it is ill suited to the techniques available.

A modified view is that object motion provides good image cues for segmentation, much as color might. This approach leads to the use of motion for segmentation, so that motion gets a more basic role in the understanding process. In this view, motion as such is useful for basic image understanding; a motion image sequence may actually be easier to understand than a static image, because the effects of motion can help in segmentation. Recent examples may be found in [Snyder 1981].

A further departure from the "static is basic" view is that motion understanding is qualitatively different from static vision. A logical extreme of this view is that there are many visual processing operations whose primitives are points in motion, and that in fact static vision is the puzzle, being ill-suited to the needs and mechanisms of biological systems. Serious work in computer motion understand-

ing has begun even more recently than computer vision as a whole, and it is too early to dismiss any approach out of hand. There are domains and applications in which the "static is basic" paradigm seems natural, but it also seems very reasonable that animals have perceptual systems or subsystems for which "motion is basic."

Section 7.2 is concerned with processing and understanding the "flow" of the world image across the retina. Section 7.3 considers several techniques for understanding sequences of static images.

7.1.1 Domain Independent Understanding

Domain independent motion processing extracts information from time-varying images using the weakest possible assumptions about the world. Processing that merely transforms the input data into another image-like structure is in the province of generalized image processing. However, if the motion processing aggregates spatial information on the basis of a common feature, then the processing is a form of segmentation.

The basic visual input for domain-independent work in motion vision understanding is *optical flow.* Although Helmholtz noted the striking immediacy of three-dimensional perception mediated through motion [Helmholtz 1925], Gibson is usually credited with pioneering the theory that a primary visual stimulus for motion is the flow of elements in the optic array, or pattern of luminance in the full sphere of solid angle surrounding the observer [Gibson 1950, 1957, 1965, 1966]. Human beings undoubtedly are sensitive to optical flow, as evidenced by the "looming" reflex [Schiff 1965], the effect of flow on balance [Lee and Lishman 1975], and many other documented phenomena [Nakayama and Loomis 1974]. The basic input to an "optical flow understander" is a continuously changing visual field, which may be considered a field of vectors, each expressing the instantaneous change of position on the optic array of the image of a world point. A field of such vectors is shown in Fig. 7.1. The extraction of the vectors from the changing image is a low-level operation often posited by optical flow research; one computational mechanism was given in Chapter 3. Flow may also be approximated in an image sequence by matching and difference operations (Section 7.3.1).

Computer vision researchers have recently begun to concern themselves with both the geometry and computational mechanisms that might be useful in the understanding of optical flow [Horn and Schunck 1980; Clocksin 1980; Prager 1979; Prazdny 1979; Lawton 1981]. Many formalisms are in use. Cartesian, polar space, and spherical coordinates all have their appeal in different situations; differential vector geometry and simple analytic geometry are both used; even the geometry of the eye or camera varies from one study to another. This chapter does not contain a "unified flow theory;" instead it briefly describes several approaches, each of which uses a different aspect of optical flow.

7.1.2 Domain Dependent Understanding

The use of models, or at least stronger assumptions about the world, is complementary to domain-independent processing. The changing image, or even the field of optical flow, can be treated as input to a model-driven vision process whose goal

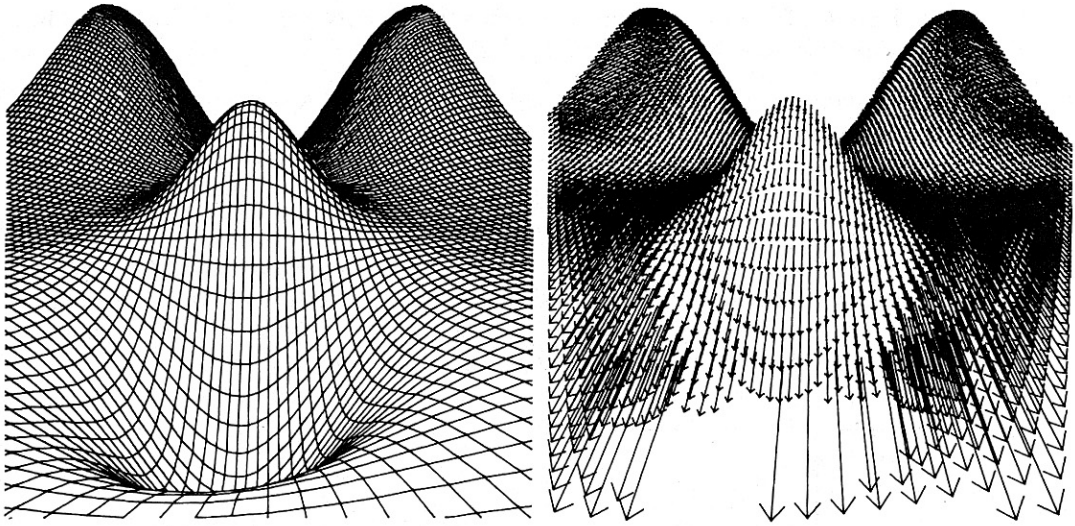

Fig. 7.1 An example of an optical flow field for an approaching "hill." (a) The hill. (b) Flow field.

is typically to segment the input into areas corresponding to meaningful world objects. The optical flow field becomes just another component of the generalized image, together with intensity, texture, or color. Motion often reveals information similar to that from range data; flow and range are discontinuous at object boundaries, surface orientation may be derived, and so forth. Object (or world) motions determine image (or retinal) motions; we shall be explicit about which motion we mean when confusion can occur.

Section 7.3 describes how knowledge of object motion phenomena can help in segmenting the flow field. One useful assumption is that the world contains rigid bodies. Tests for rigid bodies and calculations using data from them are quite useful—for example, the three-dimensional position of four points on a rigid object may be determined uniquely from three views (Section 7.3.2). A weaker object model, that they are assemblies of compound rigid pendula (linkages), is enough to accomplish successful segmentation of very sparse motion input which consists only of images of the end points of links (Section 7.3.3). Section 7.3.4 describes work with a highly specific and detailed model which is used in several ways to restrict low-level image processing and aid in three-dimensional interpretation of human motion images. Section 7.3.5 considers the processing of sequences of segmented images.

The coherence of most three-dimensional objects and their continuity through time are two general principles which, although occasionally violated, guide many segmentation and point-matching heuristics. The assumed correspondence of regions in images with objects is one example. Motion images provide another example; object coherence implies the likelihood of many "continuity" (actually similarity) conditions on the positions and velocities of neighboring image points.

Here are five heuristics for use in matching points from images separated by a small time interval [Prager 1979] (Fig. 7.2).

1. *Maximum velocity.* If a world point is known to have a maximum velocity V with respect to a stationary imaging device, then it can move at most $V\,dt$ between two images made dt time units apart. Thus given the location of the point in one image (and some assumptions about depth), this constraint limits where the point can appear on the second image.
2. *Small velocity change.* Since most visible physical objects have finite mass, this heuristic is a conseqence of physical laws and the assumption of a "small interval" between images. Of course, the definition of "small interval" depends on the definition of the velocity changes one desires to measure.

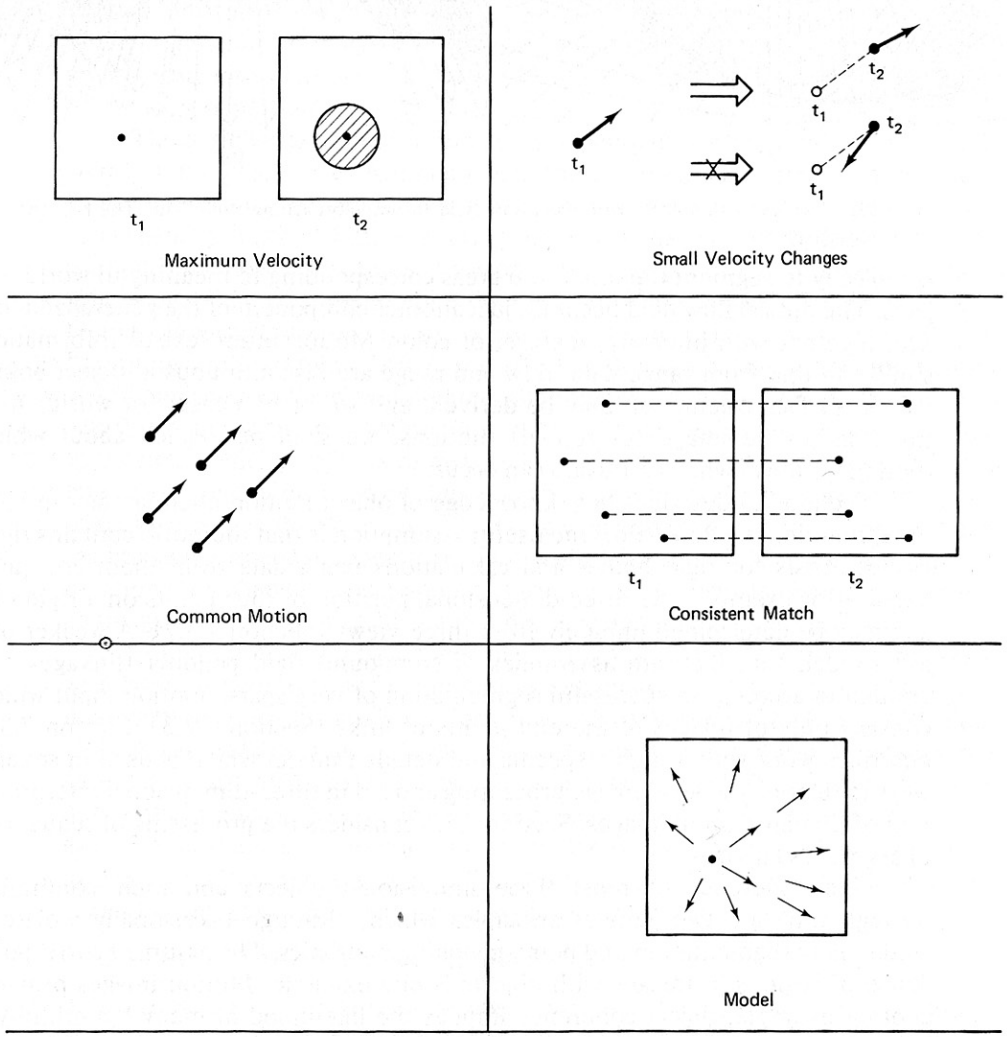

Fig. 7.2 Five heuristics.

3. *Common motion.* Spatially coherent objects often appear in successive images as regions of points sharing a "common motion." It is interesting that such a weak notion as common motion (and the related "common position") actually can serve to segment very sparse scenes of a few points with very complex motion behavior if a long–enough sequence of images is used (Sections 7.3.3 and 7.3.4).
4. *Consistent match.* Two points from one image generally do not match a single point from another image (exceptions arise from occlusions). This is one of the main heuristics in the stereopsis algorithm described in Chapter 3.
5. *Known motion.* If a world model can supply information about object motions, perhaps retinal motions can be derived, predicted, and recognized.

In the discussions to follow these heuristics (and others) are often used or implicitly taken as principles. A careful catalog of the probable behavior of objects in motion is often a useful practical adjunct to a mathematical treatment. The mathematics itself must be based on a set of assumptions, and often these are closely related to the phenomenological heuristics noted above.

7.2 UNDERSTANDING OPTICAL FLOW

This section describes some more direct calculations on optical flow, using no other input information. Information may be obtained from flow that seems useful both for survival in the world and (on a less existential level) for automated image understanding. As with shape from shading research (Chapter 3), the paradigm here is often to see mathematically what information resides in the input and to use this to suggest mechanisms for doing the computation. The flow input is assumed to be known (Chapter 3 showed how to derive optical flow by local analysis of changing intensity in the image).

7.2.1 Focus of Expansion

As one moves through a world of static objects, the visual world as projected on the retina seems to flow past. In fact, for a given direction of translatory motion and direction of gaze, the world seems to be flowing out of one particular retinal point, the *focus of expansion* (FOE). Each direction of motion and gaze induces a unique FOE, which may be a point at infinity if the motion is parallel to the retinal (image) plane.

These aspects of optical flow have been studied by computing the simulated flow pattern an observer would see while moving through a "forest" of vertical cylinders [Prager 1979] or Gaussian hills and valleys [Lawton 1981]. Some sample FOEs are shown in Fig. 7.3. Figure 7.3c shows a second FOE when the field of view contains an object which is itself in motion.

Our first model of the imaging situation is a simplification of the imaging geometry given in Appendix 1. Let the viewpoint be at the origin with the view

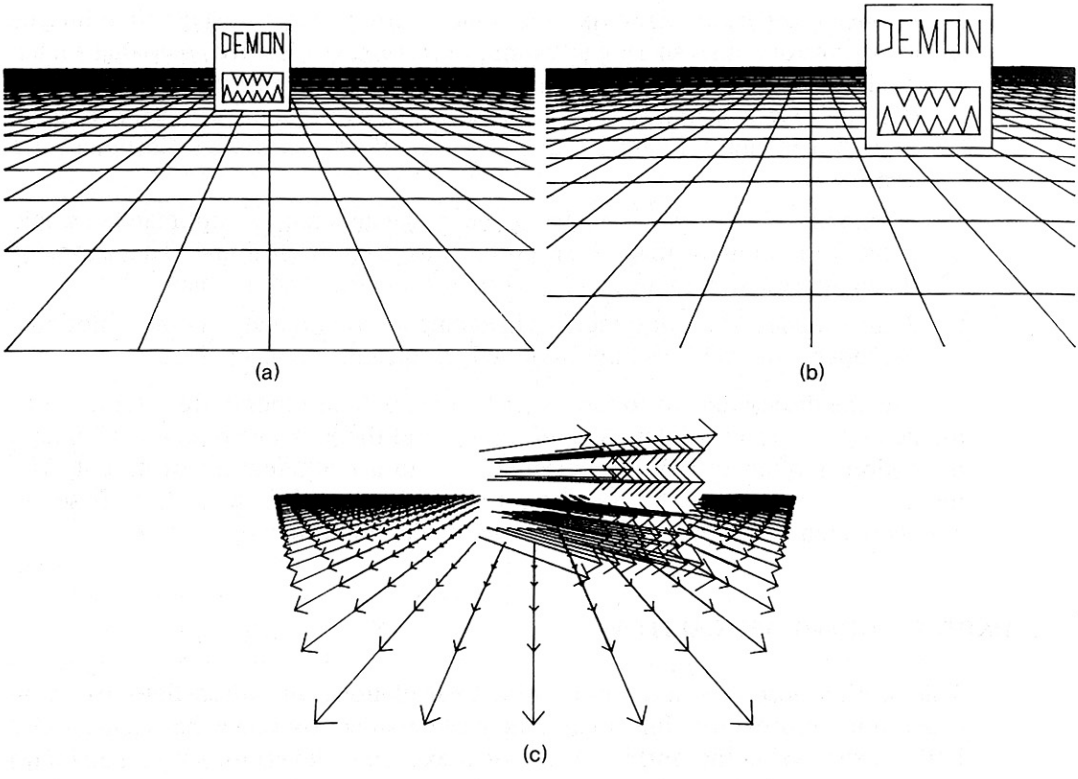

Fig. 7.3 FOE for rectilinear observer motion. (a) An image. (b) Later image. (c) Flow shows different FOEs for static floor and moving object.

direction out along the positive Z axis, and let the focal length $f = 1$. Then the perspective distortion equations simplify to

$$x' = \frac{x}{z} \tag{7.1}$$

$$y' = \frac{y}{z} \tag{7.2}$$

In the next two sections the letters u, v, and w (sometimes written as functions of t) denote world point velocity components, or the time derivatives of world coordinates (x, y, z). Observer motion with instantaneous velocity $(-dx/dt, -dy/dt, -dz/dt) = (-u, -v, -w)$, keeping the coordinate system attached to the viewpoint, gives points in a stationary world a relative velocity (u, v, w). Consider a point located at (x_0, y_0, z_0) at some initial time. After a time interval t, its image will be at

$$(x', y') = \left(\frac{x_0 + ut}{z_0 + wt}, \frac{y_0 + vt}{z_0 + wt} \right) \tag{7.3}$$

As t varies, this parametric "flow-path" equation is that of a straight line; as t goes to minus infinity, the image of the point travels back along the straight line toward a particular point on the image, namely,

$$\text{FOE} = \left(\frac{u}{w}, \frac{v}{w}\right) \tag{7.4}$$

This focus of expansion is where the optical flow originates on the image. If the observer changes direction (or objects in the world change their direction), the FOE changes as well.

7.2.2 Adjacency, Depth, and Collision

The flow path equation of a point moving with a constant velocity reveals information about its depth in z. The information is not provided directly, since all flow paths for points at a given depth do not look alike. However, there is the elegant relation

$$\frac{D(t)}{V(t)} = \frac{z(t)}{w(t)} \tag{7.5}$$

Here again w is dz/dt, and V is dD/dt. D is the distance along the straight flow path from the FOE to the image of the point. Thus the distance/velocity ratio of the point's image is the same as the distance/velocity ratio of the world point. This result is basic, but perhaps not immediately obvious.

The above relation is called the time-to-adjacency relation, because the right-hand side, z/w, is the z-distance of the point from the image plane divided by its velocity toward the plane. It is thus the time until the point passes through the image plane. This basic time interval is clearly useful when dealing with world objects; it changes when the magnitude of the world point's velocity (or the observer's) changes.

Knowing the depth of any point determines the depth of all others of the same velocity w, for it follows from the two time to adjacency equations of the points that

$$z_2(t) = \frac{z_1(t) D_2(t) V_1(t)}{V_2(t) D_1(t)} \tag{7.6}$$

The time-to-adjacency equation allows easy determination of the world coordinates of a point, scaled by its z velocity. If the observer is mobile and in control of his own velocity, and if the world is stationary, such scaled coordinates may be useful. Using the perspective distortion equations,

$$z(t) = \frac{w(t) D(t)}{V(t)} \tag{7.7}$$

$$y(t) = \frac{y'(t) w(t) D(t)}{V(t)} \tag{7.8}$$

$$x(t) = \frac{x'(t) w(t) D(t)}{V(t)} \tag{7.9}$$

As a last example, let us relate optical flow to the sensing of impending collisions with world objects. The focal point of the imaging system, or origin of coordinates, is at any instant headed "toward the focus of expansion," whose image coordinates are $(u/w, v/w)$. It is thus traveling in the direction

$$\mathbf{O} = (\frac{u}{w}, \frac{v}{w}, 1) \tag{7.10}$$

and is following at any instant a path in the environment instantaneously defined by the parametric equation

$$(x, y, z) = t\mathbf{O} = t(\frac{u}{w}, \frac{v}{w}, 1) \tag{7.11}$$

where t acts like a real scalar measure of time. Given this vector expression for the path of the observer, one can apply well-known vector formulas from analytic solid geometry to derive useful information about the relation of this path to world points, which are also vectors.

For example, the position $\mathbf{P}$ along the observer's path at which a world point approaches closest is given by

$$\mathbf{P} = \frac{\mathbf{O}(\mathbf{O} \cdot \mathbf{x})}{(\mathbf{O} \cdot \mathbf{O})} \tag{7.12}$$

where $\mathbf{O}$ is the direction of observer motion and $\mathbf{x}$ the position of the world point. Here the period (.) is the dot product operator. The squared distance Q^2 between the observer and the world point at closest approach is then

$$Q^2 = (\mathbf{x} \cdot \mathbf{x}) - (\mathbf{x} \cdot \mathbf{O})^2 / (\mathbf{O} \cdot \mathbf{O}) \tag{7.13}$$

7.2.3 Surface Orientation and Edge Detection

It is possible to derive surface orientation and to characterize certain types of surface discontinuities (edges) by their motion. A formalism, computer program, and biologically motivated computational mechanism for these calculations was developed in [Clocksin 1980].

This section outlines mainly the surface orientation aspect of this work. As usual, the model is for a monocular observer, whose focal point is the origin of coordinates. An unusual feature of the model is that the observer has a spherical retina. The world is thus projected onto an "image unit sphere" instead of an image plane. World points and surface orientation are represented in an observer-centered Cartesian coordinate system. The image sphere has a spherical coordinate system which may be considered as "longitude" θ and "latitude" ϕ. These coordinates bear no relation to the orientation of the retina. World points are then determined by their image coordinates and a range r. An observer-centered Cartesian coordinate system is also useful; it is related to the sphere as shown in Fig. 7.4, and by the transformations given in Appendix 1.

The flow of the image of a freely moving world point may be found through the following derivation. As before, let the world velocity of the point (possibly induced by observer motion) $(dx/dt, dy/dt, dz/dt)$ be written (u, v, w). Similarly,

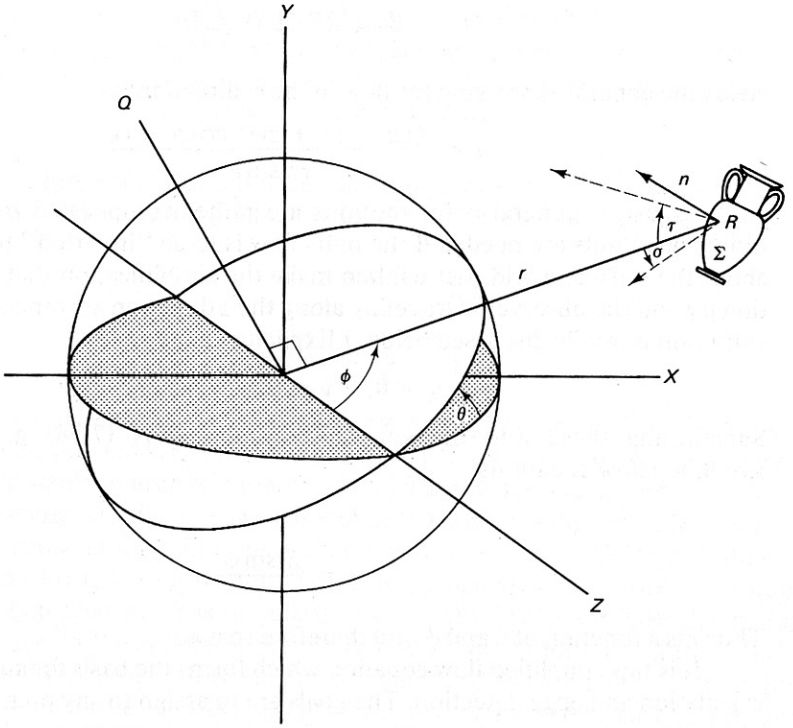

Fig. 7.4 Spherical coordinate system, and the definition of σ and τ.

write the angular velocities of the image point in the θ and ϕ directions as

$$\delta = \frac{d\theta}{dt} \tag{7.14}$$

$$\epsilon = \frac{d\phi}{dt} \tag{7.15}$$

Then from the coordinate transformation equations of Appendix 1,

$$y = x \tan\theta \tag{7.16}$$

Differentiating and solving for $d\theta/dt$ (written as δ) gives

$$\delta = \frac{v - u \tan\theta}{x \sec^2\theta} \tag{7.17}$$

Substituting for x its spherical coordinate expression $r \sin\phi \cos\theta$ and simplifying yields the general expression for flow in the θ direction:

$$\delta = \frac{v \cos\theta - u \sin\theta}{r \sin\phi} \tag{7.18}$$

The derivation of ϵ proceeds from the coordinate transformation equation

$$z = r \cos\phi \tag{7.19}$$

Differentiating, solving for $d\phi/dt$ (written as ϵ), and using

$$\frac{dr}{dt} = \frac{xu + yv + zw}{r} \tag{7.20}$$

yields the general expression for flow in the ϕ direction:

$$\epsilon = \frac{(xu + yv + zw)\cos\phi - rw}{r^2 \sin\phi} \tag{7.21}$$

As usual, general point motions are rather complicated to deal with, and more constraints are needed if the optic flow is to be "inverted" to discover much about the outside world. Let us then make the simplification that the world is stationary and the observer is traveling along the z direction at some speed S (This assumption is briefly discussed below.) Explicitly, suppose that

$$u = 0, \quad v = 0, \quad w = -S$$

Substituting these into the general flow equations (7.18) and (7.21) yields simplified flow equations:

$$\delta = 0 \tag{7.22}$$

$$\epsilon = \frac{S \sin\phi}{r} \tag{7.23}$$

Thus r is a function of θ and ϕ and therefore so is ϵ.

It is this simplified flow equation which forms the basis for surface orientation calculation and edge detection. The goals are to assign to any point in the flow field one of three interpretations: *edge*, *surface*, or *space* and also to derive the type of edge and the orientation of the surface.

To find surface orientation, represent the surface normal of a surface Σ by two angles σ and τ defined as in Fig. 7.4 with the two planes of σ and τ being the RZ and QR planes, respectively. The slant is measured relative to the line of sight, denoted by R in the figure. σ and τ correspond to depth changes in "depth profiles" oriented along lines of constant θ and ϕ, respectively. Thus,

$$\tan\sigma = \left(\frac{1}{r}\right)\frac{\partial r}{\partial \phi} \tag{7.24}$$

$$\tan\tau = \left(\frac{1}{r}\right)\frac{\partial r}{\partial \theta} \tag{7.25}$$

Surface orientation is defined by σ and τ or equivalently by their tangents. A surface perpendicular to the line of sight has $\sigma = \tau = 0$.

Equations (7.24) and (7.25) assume the range r is known. However, one can determine them without knowing r through the simplified flow equation, Eq. (7.23). The latter may be written

$$r = \frac{S \sin\phi}{\epsilon(\theta, \phi)}$$

where $\epsilon(\theta, \phi)$ gives the flow in the ϕ direction. Differentiating this with respect to θ and ϕ gives

$$\frac{\partial r}{\partial \phi} = S\,\frac{\epsilon\cos\phi - \sin\phi\,(\partial\epsilon/\partial\phi)}{\epsilon^2} \tag{7.26}$$

$$\frac{\partial r}{\partial \theta} = -\,\frac{S\sin\phi\,(\partial\epsilon/\partial\theta)}{\epsilon^2} \tag{7.27}$$

These last three equations may be substituted into Eqs. (7.24) and (7.25), and the results may then be simplified to the following surface orientation equations:

$$\tan\sigma = \cot\phi - \frac{\partial}{\partial\phi}\ln\epsilon \tag{7.28}$$

$$\tan\tau = -\frac{\partial}{\partial\theta}(\ln\epsilon) \tag{7.29}$$

These tangents are thus easily computed from optical flow. The result does not depend on velocity, and no depth scaling is required. In fact, absolute depth is not computable unless we know more, such as the observer speed.

Turning briefly to edge perception: Although physical edges are a depth phenomenon, in flow they are mirrored by ϵ, the flow measure that allows determination of orientation without depth. In particular, it is possible to demonstrate that the Laplacian of ϵ has singularities where the Laplacian of depth has singularities. An arc on the sphere projects out onto a "depth profile" in the world, along which depth may vary. If the arc is parameterized by α, relations among the depth profile, flow profile, and the singularities in flow are shown in Fig. 7.5. Thus the Laplacian of ϵ provides information about edge type but not about edge depth.

The formal derivations are at an end. Implementing them in a computer program or in a biological system requires solutions to several technical problems. More details on the implementation of this model on a computer and a possible

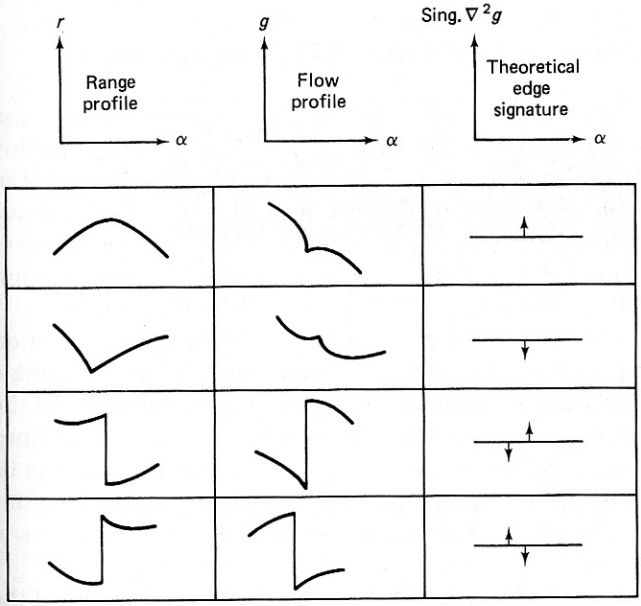

Fig. 7.5 The singularities of the second derivative of the flow profile inform about the type of edge.

implementation using low-level physiological vision primitives appear in [Clocksin 1980]. There are some data on human performance for the types of tasks attempted by the program. The assumption of a fixed environment basically implies that flow motions in the environment are likely to be interpreted as observer motions. This view is rather strikingly borne out by "swaying room" experiments [Lee and Lishman 1975], in which a subject stands in a swayable visual environment. (A large, low-mass bottomless box suspended from above may be lowered around the subject, giving him a room-like visual environment.) When the hanging "room" is made to sway, the subject inside tends to lose balance. Further, moving surfaces in the real world are quite often objects of interest, such as animals.

A survey of depth perception experiments [Braunstein 1976] points to motion as the dominant indicator of surface orientation perception. Random-dot displays of monocular flow patterns [Rogers and Graham 1979] evoke striking perceptions of solid oriented surfaces; flow may be adequate for shape and depth perception even with no other depth information. The experiments on perception of "edges," or discontinuities in flow caused by discontinuities in depth of textured surfaces, are less common. However, there have been enough to provide some confirmation of the model.

The computational model is consistent with and has correctly predicted psychological data on human thresholds for slant and edge perception in optical flow fields. (The thresholds are on the amount of slant to the surface and the depth difference of the edge sides.) The computational model can be used to determine range, but only to poor accuracy; this happens to correspond with the human trait that orientation is much more accurately determined by flow than is range. Quantitatively, the accuracy of orientation and range determinations are the same for the model and for human beings under similar conditions.

7.2.4 Egomotion

It is possible to extract information about complex observer motions from optical flow, although at considerable computational cost. In one formulation [Prazdny 1979], a model observer is allowed to follow any space curve in an environment of stationary objects, while at the same time turning its head. It is possible to derive formulae that determine the observer's instantaneous velocity vector and head rotational vector from a small number (six) of flow vectors in the image on a (standard flat) retina.

The equations that describe flow given observer motion and head rotation can be quite compactly written by using vector operators and a polar coordinate system (similar to that of the last section). The inherent elegance and power of the vector operations is well displayed in these calculations. Inverting the equations results in a system of three cubic equations of 20 terms each. Such a system can be solved by normal methods for simultaneous nonlinear equations, but the solutions tend to be relatively sensitive to noise. In the noise-free case, the method seems to perform quite adequately.

The calculation yields a method for deriving relative depth, or the ratio of the

distances of points from the observer. An approximation to surface orientation may be obtained using several relative depth measurements in a small area and assuming that the surface normal varies slowly in tne area.

7.3 UNDERSTANDING IMAGE SEQUENCES

An image sequence is an ordered set of images. The image sequences of interest here are samplings of four-dimensional space-time. Commonly, as in a movie, the images are two-dimensional projections of a three-dimensional physical world, sequenced through time. Sometimes the sequence consists of two-dimensional images of essentially two-dimensional slices of the three-dimensional world, sequenced through the third spatial dimension. Some of the techniques in this section are useful in interpreting the three-dimensional nature of objects from such spatial image sequences, but the main concern here is with temporal image sequences. In many practical applications, the input must be such a sequence, and continuous motion must be inferred from discrete location differences of image points. The thrust of work under these assumptions is often to extend static image understanding by making models that incorporate or explain objects in motion, extending segmentation to work across time [Thompson 1979, Tsotsos 1980].

When asked why he was listening to a metronome ticking, Ezra Pound is said to have replied that he did not listen to the ticks, but to the "spaces between them." Like Pound, we take the ticks, or images, as given, and are really interested in what goes on "between the ticks." We usually want to determine and describe how the images are related to each other. This information must be derived from the static images, and two approaches immediately present themselves: broadly, the first is to look for differences between the images, and the second is to look for similarities.

These two approaches are complementary, and are often used together. A general paradigm for object-oriented motion analysis is the following:

1. Segment (describe) the individual images. This process may be complex, yielding a relational structure or a segmentation into regions or edges. An important special case is the one in which the description (segmentation) process is null and the description is just the image itself. For example, an initial high-level static description is impossible if motion is to be used as an aid to segmentation.
2. Compute and describe the differences or similarities between the descriptions (or undescribed images).
3. Build a description of the sequence as a whole from the single-frame primitives and descriptions of difference or similarity that are relevant to the purpose at hand.

7.3.1 Calculating Flow from Discrete Images

This method is a form of disparity calculation that is not only used for flow calculations, but may also be used for stereo matching or tracking applications. The com-