Optic flow and depth perception

WILLIAM A. SIMPSON

Department of Psychology, University of Winnipeg, 515 Portage Avenue, Winnipeg, Manitoba, Canada R3B 2E9

Received 16 December 1991; revised 18 May 1992; accepted 20 May 1992

Abstract—The field of depth recovery from optic flow has recently experienced much growth, both on the theoretical and on the empirical fronts. Unfortunately, the theoretical results are not as widely known to perception workers as they might be. This article gives a simple analysis of the information for depth present in optic flow. It also reviews the psychophysical results for depth recovery from motion. These results are discussed with reference to the theoretical analysis and to relevant computer algorithms for depth recovery.

In his tribute to Hermann von Helmholtz, James Clerk Maxwell (1877) nicely stated the plight of the perception worker:

In no department of research is the combined and concentrated light of all the sciences more necessary than in the investigation of sensation (p. 390).

Maxwell’s observation is especially true for the study of depth perception through motion. As in all other areas of perception, knowledge of the relevant physics and physiology is required in addition to the perceptual literature. In depth from motion, however, the physics is rather complicated (3-dimensional [3D] kinematics and vector analysis), our knowledge of the physiology of visual motion detection is advancing rapidly, there is a large and technically demanding computer vision literature, and the literatures for the psychophysics of motion, depth (stereopsis) and depth from motion must all be mastered. An exhaustive review of all this literature is perhaps beyond even the capabilities of a Helmholtz. I will, however, attempt to give an adequate background for the field of depth perception through motion. I will first briefly describe the phenomenon of depth perception from motion. Next I will analyze the depth information available from visual motion. The balance of the paper is concerned with various issues in the recovery of depth from motion by humans, with references to relevant computer vision work.

1. THE PHENOMENON OF DEPTH FROM MOTION

Before we consider the information present that makes it possible to get depth from motion, let us first describe the phenomenon. It is simply this: certain static scenes look flat; movement lets the viewer assign depths to the scene elements. Helmholtz (1925) gives an early statement of the effectiveness of motion as a source of depth information:

Suppose, for instance that a person is standing still in a thick woods, where it is impossible for him to distinguish, except vaguely and roughly, in the mass of foliage and branches all around him what belongs to one tree and what to another, or how far apart the separate trees are, etc. But the moment he begins to move forward,
everything disentangles itself, and immediately he gets an apperception of the material contents of the woods and their relations to each other in space, just as if he were looking at a good stereoscopic view of it (pp. 295–296).

Bourdon (1898) was one of the first to explore the perception of depth under conditions of relative motion between stimulus and observer. Since Bourdon there have been many laboratory demonstrations that a stimulus array seen as flat when stationary jumps out in depth when it is moved. Perhaps the most familiar of these demonstrations are those of Wallach and O'Connell (1953). Objects are laid out in depth on the surface of a turntable. The shadow of the rotating objects is projected onto a screen where a viewer watches the display. This method eliminates information from convergence, accommodation, and binocular disparity, yet viewers can readily reconstruct the layout of the objects.

The depth-from-motion phenomenon is analogous to Julesz's demonstrations of depth from binocular disparity. In the absence of disparity, no depth can be seen in Julesz's (1960, 1971) random-dot stereograms. In the absence of motion, no depth can be seen in Wallach and O'Connell's displays. Just as disparity is sufficient to yield depth, so is motion.

Many studies of this sort have been done, and they will be reviewed in the psychophysics section of this paper. Although Wallach and O'Connell generated their displays by moving real objects, many authors manipulate points on a cathode ray tube in any way that might generate an impression of depth. Manipulations of velocities on the face of the screen do not necessarily simulate the rigid motion of objects in 3D space. The analysis in Section 2 will be restricted to optic flow generated by rigid motion—that is, either all the environmental points are stationary and the observer is moving, or (equivalently) the observer is stationary and the points all move together as one.

Before we look at any more results from experiments on human perception of depth from motion, let us consider the information in the projection of moving points upon which this ability is based.

2. THE DEPTH INFORMATION IN OPTIC FLOW

As an observer moves relative to his environment, the rays connecting object points to the observer's vantage point move in a regular fashion. Gibson (Gibson, 1950; Gibson et al., 1955) was among the first to note the richness of information to be found in this pattern of angular velocities, which he called 'motion perspective' or simply 'flow'. The first instance of the term 'optical flow' that I can find is in Gibson (1966, p. 161). (I prefer 'optic flow' to 'optical flow' since the former is consistent with Gibson's 'optic array'.) When the optic flow is projected on the retina, the result is a field of retinal velocities called the retinal flow. The terminology in this field is not altogether standardized; some authors use 'optic flow' to refer to what I have just defined as retinal flow. However the distinction is useful.

Many sophisticated analyses of optic (or retinal) flow have been done. Important papers include those of Longuet-Higgins and Prazdny (1980), Prazdny (1983), Koenderink (1986), and Waxman and Wohn (1988); for an analysis along quite different lines, see Lappin (1990). The basic assumption of all analyses is that the observer moves relative to a rigid environment. That is, either the observer moves amongst static objects or (identically) all the objects move together as one relative to the
Figure 1. The coordinate system for the analysis of optic flow.

observer. The coordinate system used here is centred at the nodal point of the observer's eye. This choice of axes is arbitrary but convenient. The $XY$ plane is the retina of the observer. Of course the retina is not really flat unless the eye is a video camera, but the choice of a projection surface is again arbitrary (see Maybank, 1987, Section 1.4, for a proof that the retinal flow field projected on plane and sphere are equivalent in that if one is known, the other can be deduced). The $Z$-axis is the depth axis, with $Z$-values increasing as points lie farther in front of the viewer. All motions will be described in terms of these viewer-centred axes. Hence, for example, 'Y-rotation' refers to rotation about the viewer-centred $Y$-axis. Note that a rotation about an object-centred axis is not a pure rotation in terms of our viewer-centred axes.

Longuet-Higgins and Prazdny (1980), using the coordinate system illustrated in Fig. 1, derived the retinal flow resulting from observer motion. The retina is the plane $xy$ at unit distance from the origin of the world coordinates $(X, Y, Z)$. The set of angular velocities of positional vectors like $P$ constitute the optic flow field. The velocities of their projected points on the retina constitute the retinal flow. Using vector notation, the velocity of $P$ is:

$$\mathbf{V} = -T - R \times \mathbf{P},$$

where $\mathbf{V}$ is the velocity $(V_x, V_y, V_z)$ of point $P$, $T$ is the translational vector $(T_x, T_y, T_z)$, $R$ is the rotational vector $(R_x, R_y, R_z)$, and $\mathbf{P}$ is the positional vector for point $P(X, Y, Z)$. The motion of an observer relative to the environment is described by a translational and a rotational component (each of which having $X$, $Y$, and $Z$-components; see Ramsey, 1937). These rotational and translational components in the relative motion between observer and environment in turn create rotational and translational components in the optic flow and in the retinal flow. To obtain the retinal flow we will first represent (1) in component form:

$$V_x = -T_x - R_y Z + R_z Y,$$

$$V_y = -T_y - R_z X + R_x Z,$$

$$V_z = -T_z - R_x Y + R_y X.$$
The retinal position \( p \) of object point \( P \) is given by:

\[
(x, y) = (X/Z, Y/Z).
\]  

(This is simplified by the choice of the projection plane at unit distance from the origin). The instantaneous velocity \((v_x, v_y)\) of any given retinal point with coordinates \((x, y)\) is:

\[
v_x = \frac{V_z}{Z} - \frac{XV_z}{Z^2} = \left( -\frac{T_z}{Z} - R_y + yR_z \right) - x \left( -\frac{T_z}{Z} - yR_x + xR_z \right),
\]

\[
v_y = \frac{V_z}{Z} - \frac{YV_z}{Z^2} = \left( -\frac{T_z}{Z} - xR_z + R_x \right) - y \left( -\frac{T_z}{Z} - yR_x + xR_y \right),
\]

which after some rearranging yields:

\[
v_x = x \frac{T_z}{Z} - \frac{T_y}{Z} + [xyR_x - (1 + x^2)R_z + yR_z],
\]

\[
v_y = y \frac{T_z}{Z} - \frac{T_y}{Z} + [(1 + y^2)R_z - xyR_x - xR_z].
\]

The first thing to note about the retinal flow as shown in the right hand side of Eqns 6 and 7 is that it has a translational and a rotational part. Only the translational part can give any depth information. The depth of environmental points, \( Z \), occurs in conjunction with the translational part of the flow but not with the rotational part (the terms in square brackets). In a pure rotation about the eye, all points move with the same angular velocity regardless of their depths. Figures 2 and 3 make this clear. The flows in these figures were generated using Eqns 6 and 7 for two planes at different distances from the viewer. It is plain in the translational flows of Fig. 2 that the vectors at corresponding eccentricities on the left and right sides have different lengths. The difference in depth is obtainable from the difference in velocities. In the rotational flows of Fig. 3 it is equally clear that corresponding points have vectors with the same lengths. The retinal velocity depends only on the retinal position and not at all on the depth of the point.

There is a good deal of confusion in the psychophysical literature on optic flow over the terms rotation and translation. Many workers use these terms to refer to translation and rotation in the image plane. This is confusing when used in the context of 3D motion because a translation in the world need not yield a retinal flow consisting of parallel vectors (image translation), and a rotation about the eye need not produce a retinal flow consisting of concentric vectors (image rotation). Inspection of Figs 2 and 3 should make it clear that by translational and rotational flows I do not mean 2D translations and rotations in the image plane. This usage is standard in computer vision and mathematical analyses of optic flow; I hope that it will also become standard in the psychophysical literature dealing with 3D motion.

This is perhaps a good place to clear up a long-standing misconception about the retinal flow arising from observer \( XY \)-translation. Such a flow is commonly labelled motion parallax, a term that makes obvious the relation between depth from motion and depth from binocular disparity (binocular parallax). Helmholtz (1925) pointed out the formal similarity of depth from motion and stereopsis: "[in stereopsis] the objects . . . are seen from two slightly different points of view, and, consequently there
Figure 2. Retinal flows (projections of the optic flow onto a planar retina) for pure translations of two planar surfaces. The direction is along (a) the X-axis, (b) the Y-axis, and (c) the Z-axis.

is the same kind of difference in the images as would be produced by moving in space from one place to the other” (p. 295). He also gave a crude statement of the relation between optic flow and depth: “the apparent angular velocities of objects in the field of view will be inversely proportional to their real distances away” (p. 295). If by ‘apparent angular velocities’ Helmholtz means retinal velocities, his statement is not strictly true. It is true only of pure XY-translations. If there is a Z-translational component, then the length of the vector depends on the retinal position. For example, the retinal vector has length zero at the point on the image normal to the instantaneous heading—a point known as the focus of expansion (cf Fig. 2c).

Although Helmholtz’s statement is correct for a pure XY-translational flow, it is quite misleading because even if the optic flow arises from a pure XY-translation, the retinal flow will almost always contain a rotational component. The rotational
Figure 3. Retinal flows (projections of the optic flow onto a planar retina) for pure rotations of two planar surfaces. The rotation is about (a) the X-axis, (b) the Y-axis, and (c) the Z-axis.

component arises because a viewer, when presented with a pure $XY$-translation, will not let his eye stare into space but will track objects as he moves past them. This type of involuntary eye movement is called optokinetic nystagmus (Howard, 1982).

Most textbook treatments of motion parallax depict the retinal flow as being something like Fig. 4a. This is a pure $X$-translational field for a planar surface receding in depth. In fact, after adding the eye’s $Y$-rotation, the true retinal flow will be as shown in Fig. 4b. Clearly the retinal flow is more complicated than Helmholtz’s comments suggest.

As we just have seen, analyses of optic flow often ignore the rotational part altogether (Gibson et al., 1955; Lee, 1974; Koenderink and van Doorn, 1976; Clocksin, 1980). To simplify matters, let us consider the retinal flow resulting from pure $X$-,
Figure 4. (a) The retinal flow resulting from a pure $X$-translation of a planar surface receding in depth (motion parallax). (b) The total retinal flow resulting when viewing a pure $X$-translation with tracking eye movements (an added $Y$-rotation).

$Y$-, and $Z$-translations in turn. For a pure $X$-translation:

$$v_x = \frac{-T_x}{Z}, \quad v_y = 0.$$  \hfill (8)

Pure $Y$-translation:

$$v_x = 0, \quad v_y = -\frac{T_y}{Z}.$$  \hfill (9)

Pure $Z$-translation:

$$v_x = \alpha\frac{T_z}{Z}, \quad v_y = \gamma\frac{T_z}{Z}.$$  \hfill (10)

Note that only the scaled depth is available from the retinal flow arising from a pure translation: depth is given in a ratio with velocity. This is an important point. It means that a given flow could have been produced by moving quickly relative to distant
points or slowly relative to close points. This is what allows science fiction film-makers to produce realistic spaceship fly-bys with scale models.

The relative depth of two objects can be readily obtained from a pure translational flow. In order to get relative depth (the ratio $Z_2/Z_1$) the assumption must be made that the motion is rigid. In other words, all objects must have the same velocity relative to the observer. Thus in the following equations no subscript is necessary for $T_x$, $T_y$, or $T_z$ (since these are the same for all objects). For $X$-translation,

$$v_{x1} = -T_x/Z_1, \quad v_{x2} = -T_x/Z_2,$$

The derivation is similar for a pure $Y$-translation:

$$v_{y1} = -T_y/Z_1, \quad v_{y2} = -T_y/Z_2,$$

It is particularly easy to obtain the relative depths of environmental points when the retinal flow is a pure $X$- or $Y$-translation: the ratio of the depths is given by the ratio of the retinal velocities. The situation for a pure $Z$-translation is not quite so simple; retinal coordinates must be used in addition to the retinal velocities:

$$v_{x1} = x_1(T_z/Z_1), \quad v_{x2} = x_2(T_z/Z_2),$$

$$v_{x1}/v_{x2} = x_1T_z/Z_1 \quad x_2T_z/Z_1,$$

$$v_{y1} = y_1(T_z/Z_1), \quad v_{y2} = y_2(T_z/Z_2),$$

$$v_{y1}/v_{y2} = y_1T_z/Z_1 \quad y_2T_z/Z_1.$$

Although Eqn (18) is less convenient than the corresponding equations for pure $X$- and $Y$-translation, it turns out that the depth of world points in the case of $Z$-translation is available in a particularly useful form: the time-to-collision (TTC). The TTC is the time it will take for the object to hit the image plane of the observer if the relative velocity remains constant (Lee, 1980; Tresilian, 1990). The TTC could be useful in guiding locomotion (and there is some evidence for its use by humans and animals), since it gives the absolute depth of the object in temporal terms. The TTC is defined as:

$$\text{TTC} = Z/T_z$$

where $Z$ is the distance of the object from the observer and $T_z$ is the $Z$-velocity of the object. The TTC gives absolute depth information: the object will hit in so many seconds. This sort of information is different from the relative depth information discussed earlier: object 1 is so many times further away than object 2.

The TTC is available from retinal quantities. In fact it is simply the ratio of the retinal coordinate to its velocity:

$$x/v_x = x/T_z/Z = Z/T_z = \text{TTC},$$
Figure 5. (a) A compound retinal flow resulting from $X$-translation past two planar surfaces with rotations about the $X$-, $Y$-, and $Z$-axes. (b) The rotational component of the retinal flow. (c) The translational component of the retinal flow.

$$\frac{y}{v_y} = \frac{y}{yT_z/Z} = \frac{Z}{T_z} = \text{TTC}.$$ (21)

The TTC could be particularly useful in regulating action. For example, a batter has to ensure that the bat meets the ball at a particular time and place. If he can determine visually when the ball will be in striking range, all he has to do is make sure that his bat is there at the same time.

In our derivation of relative depth from the retinal flow we have made the simplifying assumption that the flow has no rotational component. If we were to attempt to get relative depth from a compound flow (one with both translational and rotational components such as that shown in Fig. 5a), the results would not be correct. The
computations of relative depth must be obtained using only the translational component (Fig. 5c). Any process or mechanism for getting depth from optic flow must somehow filter out the rotational component (Fig. 5b). In fact, this is the problem for any system trying to recover depth from motion (we will treat the problem fully in Section 3.2). Once this filtering has been achieved, it is trivial to obtain the depth.

2.1. Derivatives of depth from derivatives of optic flow

We have just seen how relative depth can be obtained from the translational component of retinal flow. However more than just the raw depths of world points can be obtained if we assume that the depths of the points change smoothly (i.e. the points are part of a smooth surface). Longuet-Higgins and Prazdny (1980) have shown that the gradient of the surface (its slant and tilt) at a point can be recovered from the optic flow.

Just as the slant and tilt of a surface are found by taking spatial derivatives of the surface coordinates, so is the gradient of a surface recovered from optic flow by taking spatial derivatives of retinal velocity vectors. When we differentiate the retinal vectors we find the rate at which they change in the x- and y-directions. If we represent $u_y$ by $v_x$, the four spatial derivatives are: $u_x$, $u_y$, $v_x$, $v_y$. These derivatives of the retinal velocities can then be used to recover the gradient of the surface. (In fact Longuet-Higgins and Prazdny also use the spatial second derivatives of the retinal velocities.)

Longuet-Higgins and Prazdny point out that certain combinations of the retinal velocity derivatives would be useful. Networks of motion sensors that could compute these combinations are illustrated in Fig. 6. Adding $u_y$ and $v_y$ gives the divergence or div. The divergence of a 2D vector field is a scalar field that tells us, at each point, how much the field diverges (how much the image expands) away from that point. The difference of $u_x$ and $v_x$ is the vorticity or curl. The curl of a 2D vector field is a scalar field that gives, at each point, an indication of how the field swirls (how much the image Z-rotates) in the vicinity of that point. Finally, the combinations $u_x - v_y$ and $u_y + v_x$ are the two components of shear or deformation. The deformation (def) is a differential operator yielding a scalar field that tells how much the field contracts in one direction and simultaneously expands in an orthogonal direction (there are two such operators). These combinations of velocity derivatives are 'differential invariants'—they do not depend on the choice of a coordinate system (Koenderink and van Doorn, 1976, 1981; Koenderink, 1985, 1986). This clearly is an advantage for those wishing to model biologically plausible mechanisms sensitive to optic flow.

Recovering the orientation of a surface at a point is not really useful unless the surface is planar. Subbarao (1988) shows that if the surface is better approximated by a curved patch, the curvature can be obtained from second spatial derivatives of the retinal velocities.

To summarize, the following information about the depth of world points can be recovered from optic flow: (1) The relative depths of points can be recovered from the translational retinal flow. (2) The orientation of the surface and its curvature can be obtained by using first and second spatial derivatives of the retinal velocities.

3. HUMAN EXTRACTION OF DEPTH FROM OPTIC FLOW

Although the most logical way to go about studying a perceptual area is to first
analyze the information present in the stimulus and then see if the organism is sensitive to that information, empirical research on structure from motion long preceded theoretical analysis. The past research in this field has been largely phenomenon-driven. Very little in the way of theory appeared before Gibson (1950), and detailed analysis of optic flow has only recently been done (Koenderink and van Doorn, 1976, 1981; Longuet-Higgins and Prazdny, 1980; Prazdny, 1980, 1983; Rieger, 1983; Koenderink, 1985, 1986; Waxman and Wohn, 1988).

Having described the information available in optic flow, we can now sort through the psychophysical literature in a search for clues about how this depth information is extracted. There are several issues that fall under this general topic, and I will discuss relevant results from computer vision and human psychophysics.

3.1. What is the input to the depth-from-motion process?

The name 'depth-from-motion' seems to presume the answer: the input to the process that derives depth from optic flow is retinal motion. This appears moderately safe, but in fact it is a matter of some dispute. Some propose that the human ability to recover depth from motion might be based on sequential readings of retinal point position, other nonvelocity information, retinal velocities, or spatial derivatives of retinal velocities.
3.1.1. Sequential readings of retinal point position. An extreme skeptic could maintain that all performance on what are nominally motion tasks is really based on registering retinal positions at one time and comparing them to those obtained at another. Formally, of course, whenever motion is presented, so are positional cues (position is the integral of velocity). There are a few arguments that lead us to reject this extreme view.

Subjective impression of motion. The weakest argument would be that there certainly is a subjective difference between seeing a clock's second hand move and surmising that the minute hand must have moved because it is in a different position now than it was before.

Random dots reveal motion mechanism. A stronger argument comes from Nakayama and Tyler (1981). They reasoned that a person may keep track of the positions of a few points, but it would be impossible to keep track of the positions of a whole field of random dots. They obtained different sensitivities to sinusoidal motion of random-dot patterns and single lines, and argued that these different results were due to the dots being processed by a motion mechanism and the lines by a positional mechanism. However, positional cues are still there in the random dots and the lines are seen to wiggle.

Adaptation and aftereffects. The strongest psychophysical evidence for motion sensors is the presence of adaptation and aftereffects (for a review see Anstis, 1986). After staring at a moving pattern for some time, sensitivity to motion in that direction declines (adaptation). If the pattern should then stop, it will appear to move in a direction opposite to the adapting direction (after-effect). It is hard to see how these effects would be explained by a simple positional mechanism.

Lesions produce motion-specific deficits. Neurological evidence from Zihl et al. (1983) is perhaps the most convincing. The authors did extensive tests on a patient with brain damage in the lateral temporal-occipital region. This area is near the medial temporal area which is known from electrophysiology to be specialized for motion detection (Maunsell and van Essen, 1983; Albright, 1984; Newsome et al., 1989). The patient's spatial (acuity) and temporal (critical flicker frequency) abilities were normal, but her motion perception was greatly impaired. For example, she reported seeing a target moving down about 55% of the time compared to 100% for a normal subject. The patient reported difficulties in such everyday activities as pouring a cup of tea because of her inability to see motion (the tea appeared to be frozen; she could not tell when to stop pouring because she could not see the rising level of the tea in the cup).

So the evidence for motion sensors in the human visual system is quite good. It still remains to be seen, though, whether it is their output that feeds the depth-from-motion process. This is not necessary from an algorithmic view. Several computer vision approaches treat the input as a sequential series of snapshots (e.g., Ullman, 1979a,b; Longuet-Higgins, 1981; Tsai and Huang, 1984). From the viewpoint of experimental methodology, human performance on the depth task might really be based on other cues.

Sperling and associates (Sperling et al., 1989, 1990) argue that subjects' performance in many previous studies of depth recovery from displays of objects rotating
about a self-centred Y-axis (the ‘kinetic depth effect’) could have been based on shortcut computations. Some of these possible shortcuts include deriving at least some retinal velocities, but others do not. Dot density is an example of a nonmotion cue that could be used. If the dots on a surface are evenly distributed, their projection onto the retina does not consist of evenly distributed dots. The dot density in a single frame could possibly be used to recover the shape of the object. Sperling et al. (1989) found that dot density was in fact a weak cue to depth. The general argument remains, however.

A prime example of a depth-from-motion study where performance could be based on positional cues is from Graham et al. (1948). The subjects had a monocular view of two needles that moved rigidly sideways together. The task was to adjust one needle until it was seen to be in the same plane as the other. It is not necessary to see any depth in the display to do this task; the subject can simply equate the needles’ retinal velocities. With these stimuli, it is not even necessary that velocities be used. If Nakayama and Tyler argue that positional cues are used when viewing their wiggling line stimulus, we could easily say that positional cues are used here. Through use of random-dot stimuli (as in Nakayama and Tyler) the importance of static positional cues might be reduced.

Other nonmotion cues include changes in dot density, line length, or area. Koenderink (1986, p. 169) points out that changes in texture density could be used as a measure of depth. Mechanisms sensitive to the rate of change of texture density, line length, or area over time would not necessarily be sensitive to retinal velocities. We will consider such mechanisms next.

3.1.2. Other methods not using retinal velocities. Some methods for recovering depth from time-varying imagery do not take static snapshots as input but do not use retinal velocities either. One basic approach to recovering depth without explicitly recovering motion involves computing spatial and temporal derivatives of image intensity. Buxton and Buxton (1983), Zinner (1986), and Negahdaripour and Horn (1989) take this general approach, which I will now review in more detail.

Buxton and Buxton (1983) propose two depth-from-optic-flow mechanisms for the restricted case of pure Z-translational flows. The first step is a convolution of the image intensity signal \((x, y, t)\) with a d’Alembertian of a Gaussian. Like Marr’s (1982) Laplacian of a Gaussian, this filtering operation produces edge estimates at the zero-crossings. In the first depth mechanism, Buxton and Buxton propose that the ratio of the rate of intensity change in the filtered image in a given direction to the rate of intensity change over time in the filtered image can be used to give the scaled depth directly. This first method is really just using retinal velocities under a different guise—image velocity is a ratio of temporal and spatial frequencies (see Ator, 1963). The second method again uses the filtered image, but only in the periphery. The filter produces two types of zero-crossings: static zero-crossings, which are in the position of an edge, and ‘depth zeros’ which may be used to infer the depth of the object. When these two types of zeros coincide, the depth can be readily obtained as a TTC.

Zinner (1986) models the scene as a plane (though the method can be extended to nonplanar surfaces). This would be a good approximation for the type of task his method was designed for: guiding an airborne vehicle. The ground below would be rather planar, especially if viewed from some height. Zinner combines the retinal flow
equations arising from motion relative to a plane with equations describing the effect of motion on the temporal derivative and several spatial derivatives of image intensity. The resulting set of equations can be solved to give the sensor motion and relative depth. The algorithm seems to work quite well on natural image sequences.

Negahdaripour and Horn (1989) present a method for recovering the observer motion relative to an arbitrary scene. Once these motion parameters are recovered, the depth is also available. As in Zinner’s algorithm, spatial and temporal derivatives of the image intensity are used.

Another approach to depth recovery from time-varying images without velocity extraction is to regard motion as orientation in space-time. Bolles et al. (1987) present a method for the limited case of pure translation. The translation must be along the X-axis: the camera looks to the side while moving forward. At each instant the image is a 2D array of intensities. Over time, the images are stacked, yielding a solid (x, y, t) of intensities at different locations and times. Figure 7 shows a black dot moving (in discrete steps) from the left bottom to the right top of a display. In the space-time solid, the result is a dotted line. The slope of the line reveals the speed and direction of the image feature’s motion. In Bolles et al.’s work, it is the camera that moves in a pure X-translation. The sloping lines in the space-time solid reveal the depth of the objects, since in a pure X-translation the scaled depth is given directly by the image velocity.

In the psychophysical literature Mather (1991) and Adelson and Bergen (1985) take the same approach, though they are interested only in motion, not depth-from-motion. Psychophysicists tend to reduce 2D retinal images to one-dimensional cross-sections—both authors show planes in (x, t). The basic point is the same: motion is orientation in space-time.
Figure 8. The object with length $Y$ at distance $Z$ from the viewer subtends the visual angle $\alpha$.

Changing length or area. If the observer's motion contains a $Z$-translational component, the depth can be derived by monitoring the rate of change of the length of a retinal image. The following proof is essentially that of Carel (1961). The geometry is shown in Fig. 8. An object of length $Y$ is at distance $Z$ from the observer. By simple trigonometry,

$$\tan \alpha = \frac{Y}{Z}. \quad (22)$$

If the visual angle $\alpha$ is small, $\tan \alpha$ approximately equals $\alpha$:

$$\alpha = \frac{Y}{Z}. \quad (23)$$

The object approaches the observer at velocity $T_z$:

$$Z = T_z t. \quad (24)$$

Substituting and taking the derivative of $\alpha$ we get the TTC:

$$\text{TTC} = -\frac{\alpha}{dx/dt} = -\frac{Y/T_z t}{-Y/T_z t^2}. \quad (25)$$

An organism could potentially get the TTC without using retinal velocities, but by instead monitoring the image length and rate of length change. Maybank (1987, pp. 190–201) has shown that the TTC can also be obtained from changing retinal area.

The TTC is a useful piece of information, but do organisms derive TTC optically? Several studies find evidence for sensitivity to TTC: Carel (1961), Schiff and Detwiler (1979), Lee and Reddish (1981), Todd (1981), Wagner (1982), Simpson (1988), and Wang and Frost (1992). It is possible that something other than TTC was used by the subjects in these studies. For example, Borst and Bahde (1988) present evidence suggesting that Wagner's subjects (which were flies) were responding not to TTC but to a spatially and temporally integrated signal from velocity sensors. The details of each experiment were different, however, and the replication of results in the different situations suggests that TTC is calculated visually.

It appears that visual systems are able to compute TTC. The next question is: how exactly do they do it? There are at least three possibilities: they could be using retinal velocities and positions (Eqns 20 and 21), image length and rate of length change (Eqn 25), or image area and rate of area change (Maybank, 1987). Despite the number of studies on TTC, few allow any inferences about the mechanism involved.

In Simpson (1988) I studied TCC discrimination using stimuli that were crosses of horizontal and vertical line segments. The line widths did not change, only their
lengths. Subjects were able to discriminate TTC with these stimuli, supporting the idea that the mechanism used line length and rate of line length change. The experiments also support the existence of a mechanism that uses retinal position and velocity, however. I compared the observers' ability to discriminate TTC for sequences of an approaching or withdrawing object. So far as the depth information contained in the sequences goes, the two types of display are identical. However, Ball and Sekuler (1980) found that viewers are more sensitive to retinal motions away from the fovea (such as produced by approaching objects) than to motions toward the fovea (such as produced by withdrawing objects). We would predict from this result that if the TTC mechanism uses retinal velocities (rather than line length or area change), discrimination should be impaired for withdrawing objects. This prediction was confirmed by the experiment. Perrone (1986) found a similar asymmetry for approaching and withdrawing objects and used a similar explanation. This parallel of asymmetrical ability to detect depth in displays of approaching and withdrawing objects and asymmetrical ability to detect retinal velocities suggests that retinal velocities are in fact used in computing TTC.

Regan and Beverley have done many important studies which cast light on the TTC mechanism (for a review see Regan et al., 1986). Most important for our purposes was their discovery (Regan and Beverley, 1978) of visual mechanisms sensitive to changing size. They found that adaptation to a solid square whose edges oscillated in antiphase with each other (resulting in changing size) elevated thresholds for detecting changing size, but had little effect on thresholds for detecting motion of a square whose edges oscillated in phase with each other (resulting in a diagonal motion of the square), and vice versa. Such mechanisms could be TTC ('looming' or div) detectors. The effect of adaptation was found to be localized in the retinal area adapted, leading to an estimate of the looming detectors' receptive fields as being about 1.5 deg in diameter (Beverley and Regan, 1979). Regan and Cynader (1979) found looming detector cells in cat area 18. Saito et al. (1986) also found cells (in macaque MST cortex) responsive to expansion or contraction.

Although Regan himself refers to the cue as changing size, we cannot be sure that the mechanism uses changing line length or changing area. It could use a combination of retinal velocities as shown in Fig. 6. (In fact Regan, 1986, proposed a similar model.) One factor that favours changing size as a mechanism is the relatively sparse retinal flow arising from the stimulus Regan used. The stimulus is a solid homogeneous square, so no velocities could be derived for points other than those at the edges. Saito et al. (1986) found some cells that responded better to expansion or contraction of a solid shape rather than a dot pattern. However these authors also found some cells that preferred dot patterns, and some that responded equally to the expansion/contraction of dots or solids.

To summarize, the psychophysical evidence by and large does not allow definitive statements on whether the visual system uses changing length or area to recover depth from optic flow. The best evidence for changing area comes from Regan, whose stimuli had no internal structure and thus would yield sparse retinal flows. Retinal velocities would still be available at the edges, though. The most favourable evidence for changing line length comes from Simpson. However the finding of asymmetrical perception of approaching and withdrawing objects (Simpson, Perrone) is explained by differential sensitivity to retinal velocities toward and away from the fovea. This
is evidence that retinal velocities, not changing length or size, feeds the depth-from-motion process. At this point, the case is not decisive for any mechanism. Evidence presented in Section 3.2 makes it unlikely that the visual system uses changing length or size, and more likely that it uses retinal velocities.

Rate of texture change. Koenderink (1986) has shown that the div can be obtained by monitoring the rate of change of the texture density in an image patch. Surprisingly, it is possible to obtain the combination of a pair of spatial derivatives of the image velocity without first obtaining the velocity. Does the visual system work this way?

Beverley and Regan (1983) did some experiments that bear on the question of changing texture density and depth. The authors independently varied the rate of change of a square’s size and the rate of change of the size of texture elements within the square. The squares’ interiors were random checkerboard patterns. As one might expect, best after-effects occur when there is both size and texture change. An unexpected result, though, is that after-effect magnitude increases with rate of texture element change even if it exceeds rate of size change. These experiments do not necessarily implicate a mechanism sensitive to changing texture density. The changes in the texture could have easily been detected by mechanisms monitoring the size of each interior checkerboard square, or the velocities of each interior edge. The fact that after-effects are increased when both the interior texture and the outside edges of the square change size suggests that looming sensors might really work by combining retinal velocities. A pure area-change mechanism would presumably work as well with textured as with untextured squares, so long as the outside dimensions of the square fell within its receptive field (and they did in these experiments).

The following type of stimulus is needed to demonstrate the existence of sensitivity to changing density. A random-dot covered surface approaches the observer. The dots do not change size as the surface approaches. A different set of random dots covers the surface on every frame. The display will look like random moving noise that gets less and less densely packed (this would be the 2D interpretation, at least). Sperling et al. (1989) used such a display, and subjects could not recover the depth from it. It seems that humans cannot use the changing density cue.

To summarize this section, the psychophysical evidence gives little support to the idea that the depth-from-motion process uses such nonvelocity inputs as changing length, size, or density. The most likely candidate for the input is retinal velocity.

3.1.3. Retinal velocities. We have considered several nonvelocity cues as potential inputs to the human depth-from-motion process. For various reasons, these seem untenable. It is quite likely that the process for recovering depth uses retinal velocities.

The problem of sensing retinal motion has been the subject of intense study in the last 15 years (Nakayama, 1985a). Many schemes for motion sensors exist (Reichardt, 1961; Ator, 1963; Barlow and Levick, 1965; Horn and Schunk, 1981; Marr and Ullman, 1981; van Santen and Sperling, 1984, 1985; Adelson and Bergen, 1985; Watson and Ahumada, 1985; Heeger, 1988; Bülthoff et al., 1989; Fleet and Jepson, 1989; Grzywacz and Yuille, 1990; Perrone, 1990). See Scott (1988) for a recent review of computational approaches to this problem, and for a novel algorithm. Velocity is a vector quantity having both magnitude and direction. Exactly how such a 2D velocity field (the retinal flow) is determined is beyond the scope of this paper—the
topic would require an entire review paper of its own. There is a wealth of psy-
chophysical evidence (Sekuler, 1975; Sekuler et al., 1978) that the human visual
system computes the retinal flow. Some of this evidence was reviewed in Section 3.1.1

One of the strongest arguments that retinal velocities are used to recover depth is
that the depth-from-motion mechanism fails at isoluminance. It has been shown that
the motion signal is considerably weakened when the visual contours are defined by
hue rather than luminance (Cavanagh et al., 1984). If retinal velocities are the input
to the depth recovery mechanism, observers should be unable to see depth in isolu-
minant optic flow. This was the result reported by Livingstone and Hubel (1987) in
humans and by Lehrer et al. (1988) in bees. Similarly, Rivest et al. (1990) found that
ability to see depth from motion in isoluminant displays was reduced, though not as
much as would be predicted from the loss in 2D motion sensitivity. However, the
situation is not as simple as might be hoped. Cavanagh (1990) reports that depth is
seen in an isoluminant display of an object rotating around its own Y-axis.

Dosher et al. (1989) argued that retinal velocity rather than sequential position
must be the input to the depth-from-motion process because the ability to derive
depth from motion remains unimpaired as the time a given dot remains in the image
(dot lifetime) is reduced from 20 to 2 frames. Of course, sequential positions can still
be obtained when the lifetime is 2 frames, but it is unlikely that the visual system notes
the position of each dot, especially when they are continually being replaced by new
ones. More cogent is their argument from experiments with alternating contrast and
interposed blank grey frames. Each of these manipulations is known to impair the
performance of simple motion detectors (e.g., Reichardt, 1961). The ability to recover
depth was impaired with these stimuli, suggesting that the input to the depth-from-
motion mechanism is retinal velocity as computed by simple motion sensors. Landy
et al. (1991) extended these observations, finding that the manipulations that impair
simple motion detection also impair depth recovery.

Lesion studies by Siegel and Andersen (Crawford et al., 1990, pp. 49–50) point to
velocity as the input to the depth-from-motion process. Monkeys with lesions in MT
cortex were not able to see the 3D structure in a rotating dot display, though their
performance beforehand had been similar to humans'. Performance on a 2D motion
task also declined after the MT lesions. MT cells are known to be specialized for
motion detection; since lesioning the area degraded depth-from-motion performance,
the inference is that these motion-sensing cells normally provide the motion signal
used to recover depth. However this inference is threatened by the fact that the
performance on the 2D task recovered much more quickly after the lesioning than did
performance on the 3D task.

The psychophysical evidence suggests that the input to the depth-from-motion
process is retinal velocity. How does the visual system use these retinal velocities to
recover the depth? Conveniently, we can turn to the computer vision literature for
ideas since most computer vision algorithms for extracting depth from optic flow take
the retinal flow (the 2D field of retinal velocities) as input. Generally speaking, a
retinal flow will have both a translational (expansion or shear) and a rotational (rigid
image displacement) component. Refer again to the compound flow shown in Fig. 5a.
The main problem for a depth-from-retinal-flow algorithm is how to extract the
translational component shown in Fig. 5c from the compound flow; we could
rephrase the problem as being that of how to filter out the rotational part shown in
53

Fig. 5b. Some algorithms that solve the problem of getting depth from compound flows are reviewed in Section 3.2.

3.1.4. Spatial derivatives of retinal velocities. We have seen in Section 2.1 that spatial derivatives of retinal velocities can be used to find the orientation or curvature of a surface. However, there are reasons to doubt the usefulness of such spatial derivatives in a practical vision system (cf. Longuet-Higgins, 1986).

The retinal velocities themselves are often not well-behaved. They may be undefined in many regions (where there is little visual texture), they may change discontinuously (at object boundaries), and they may even be multivalued (as when moving past a dirty window and looking through it at the scenery behind). Obviously any of these situations makes computation of spatial derivatives of the retinal velocities problematic. Furthermore, the visual system seems best able to derive depth from optic flows that are quite difficult to handle mathematically. For example, the most vivid depth impressions arise from relative motion of a background viewed through a foreground (the dirty-window case). Conversely, planes and smooth surfaces that give rise to well-behaved retinal velocities are, even on a theoretical level, hard to interpret (see Section 3.3).

The method of Longuet-Higgins and Prazdny (1980) for recovering the observer motion and orientation of a scene patch requires not only first spatial derivatives, but also second derivatives of the image velocities. Longuet-Higgins (1986, p. 182) later feared "that such a computation would be altogether too ill-conditioned to be of any practical use". There do exist algorithms (e.g., Section 4.3 of Maybank, 1987) that use only the first derivatives, but as we have seen even these may be problematic.

On a psychophysical level, it is hard to say that a practical pattern of results indicates extraction of spatial derivatives of retinal velocities rather than simple velocities. Even Regan's evidence for looming sensors, which Longuet-Higgins and Prazdny (1980) suggested might be div-detectors, is ambiguous. Clearly the output of such a sensor can be interpreted as a combination of spatial derivatives of retinal velocity, but does this clearly describe how the mechanism works? It could, for example, be using changes in length or area instead of retinal velocities. As well, Regan's results could be explained by referring simply to the retinal velocities themselves rather than to their spatial derivatives. We will restrict our discussion to the ramping or unidirectional motion condition in Regan and Beverley (1978). The main finding was that adaptation to retinal expansion elevated thresholds for subsequent detection of retinal expansion but not sideways movement. This is precisely what would be predicted if the retinal velocities themselves were at work (see Fig. 9). During adaptation to expansion, the motion sensors for each retinal direction are fatigued; the motion aftereffect is building up. The perceived motion in the test phase will be the vector sum of the aftereffect and the actual motion of the test display. If both the adapt and test motions are expansion, the motion sensor at each location is receiving the same direction of motion as in the adapt phase, and hence it responds less. If a sideways motion is viewed in the test phase, half the sensors receive the same direction of motion as in the adapt phase, but half do not. The sensors receiving the same motion direction will respond less, but those receiving the opposite direction of motion will respond more. The net effect of expansion adaptation on sideways moving
Figure 9. Observers in Regan and Beverley (1978) adapted to a square that changed size; they then viewed the same stimulus (bottom) or a square that moved diagonally in the plane (top). The threshold for detecting changing size was elevated but the detection threshold for diagonal motion was not. This result is predicted if we suppose that the neural after-effect vectors are added to the test stimulus.

test stimuli will be nil. So Regan and Beverley's results could be due to the adaptation of local retinal velocity sensors rather than to adaptation of div detectors.

Psychophysical determination of spatial contrast sensitivity functions for motion makes it unlikely that spatial derivatives of velocity are computed in human visual systems. Nakayama and Tyler (1981) showed the motion sensing system to be a lowpass spatial filter (see also Nakayama, 1985b). The display was a field of random dots whose velocity was a sinusoidal function of vertical position. If the spatial frequency of the motion rises above about 0.6 cycles/deg, detection thresholds rise. Taking spatial derivatives is the same as spatial highpass filtering; we can see that the observed behaviour is the opposite of what we would expect if the visual system were computing spatial derivatives of retinal velocities.

3.2. Dealing with the rotational component

As I have noted several times, the optic flow and the retinal flow contain both translational and rotational components. Only the translational component can give depth information. How does the visual system do this decomposition of the retinal flow? I will review some possible algorithms from the computer vision literature, whose plausibility we can evaluate in light of physiological and psychophysical evidence.

Prazdny (1981a) has outlined a method for finding the translational component that is based on the fact that the field of the retinal velocities resulting from the translational component consists of motion along straight lines all intersecting at a common point, the focus of expansion (see Fig. 2c). Prazdny's method begins by extending all retinal vectors towards the centre. If a rotational component is present, the vectors will not all intersect at one point. By minimizing the dispersion of the intersections of the velocity vectors (the dispersion is zero for the transitional field) as
a function of six parameters (the rotational and translational components), both the translational and rotational fields are found. Note that this method is designed to work with flows having some $Z$-translational component. Prazdny suggests that if a flow having no $Z$-translational component is detected the following modification can be used. Instead of minimizing the dispersion of the vectors' intersections, maximize it. If there is no $Z$-translational component, the translational flow vectors will all be parallel (cf Figs 2a and 2b).

Both Longuet-Higgins and Prazdny (1980) and Rieger and Lawton (1985) capitalize on the fact that the optic flow field (or retinal flow field) does not necessarily assign only one vector to each direction (or retinal point). Two texture elements lying in the same direction but at different depths—e.g., dirt on a dusty window and a point in the scene behind, or a point at an occluding edge and one behind it—will have the same rotational velocities, but will differ in their translational velocities (as always, we assume the optic flow arises from a rigid motion). Thus in a retinal flow field resulting from translation and rotation, the difference of vectors at a retinal point will yield the difference of their translational velocities. This in turn yields the focus of expansion, the translational field, and the depths of all points. These schemes using multi-valued optic flow are informally supported by the vividness of the depth percept in the dirty window case. Further support for the Rieger and Lawton scheme comes from its confirmed prediction that the ability to detect heading improves as the range of depths in the scene increases (Rieger and Toet, 1985; Cutting, 1986; Warren and Hannon, 1988). As with depth, the heading depends on the translational component—it is given as a normal to the focus of expansion (Fig. 2c). (The focus of expansion's retinal coordinates are $(T_x/T_z, T_y/T_z)$).

Although all the depth information is contained in the translational flow, it is not necessary from an algorithmic point of view to first remove the rotational component. We can deal with both at the same time. One way to do this would be to eliminate $Z$ from Eqns (6) and (7) and combine all terms into one nonlinear equation. We can then take several retinal points and solve the nonlinear equations for the unknown parameters. Because these simultaneous equations are nonlinear: (a) there is no analytic solution; and (b) multiple solutions are possible. This basic approach is contained in Prazdny (1980), Waxman and Ullman (1985), Horn (1986, chapter 17), and Koenderink and van Doorn (1987). The problems associated with solving nonlinear equations do not necessarily rule these schemes out as plausible models for the human depth-from-motion process. Although the iterative solution of such equations is slow on ordinary computers, it may be implemented in parallel in humans. And, as Waxman and Wohn (1988) point out, the multiple solutions may correspond to illusions and ambiguities in human perception.

Due to the difficulties inherent in dealing with nonlinear simultaneous equations, linear simultaneous equations are certainly to be preferred. Algorithms for extracting depth from retinal flow using sets of linear simultaneous equations have been proposed by Longuet-Higgins (1981), Mitiche (1988) and Zhuang et al. (1988). An excellent discussion of Longuet-Higgins's algorithm can be found in Scott (1988). Closed form (as opposed to iterative) solutions to the depth from optic flow problem have been given by Longuet-Higgins (1984) and Waxman and Wohn (1988). Both involve finding the roots of a cubic equation. Longuet-Higgins considers the case of...
a flow generated by a plane (see Scott for a lucid exposition); Waxman and Wohn extend the analysis to flows from curved surface patches.

The previous schemes took as their input the retinal flow field—the velocity at every retinal point. Koenderink (1985) and Regan (1986) have proposed a scheme where mechanisms composed of a network of motion sensors, each responding to a given direction of motion, analyze the flow. Instead of first sensing the total (translation + rotation) flow and then somehow removing the rotational component, these sensors can respond solely to the Z-translational component at the initial registration of the flow. Figure 6 shows sensors that would yield div (Z-translation, looming, or retinal expansion/contraction), curl (Z-rotation), and def fields. Koenderink analyzes the flow into div, curl, and def components.

If looming detectors as proposed by Regan and Koenderink underlie human performance, we would expect pure Z-translational depth discrimination to be unaffected by the addition of a rotational component. A rotation simply displaces the whole image. Since a looming sensor responds only to image expansion, it is able to filter out the nuisance rotational component perfectly. I found that viewers' sensitivity to depth conveyed by Z-translation decreased as I added various amounts of rotation about the observer's $X_-$, $Y_-$, and $Z_-$axes (Simpson, 1988). This suggests that the depth-from-motion mechanism uses the total flow, not just the translational component. With this assumption and given Weber's law for velocity discrimination (McKee, 1981), we would predict the degradation of discrimination performance found. If the total flow is compared for the standard and comparison object, then adding a rotation will increase the retinal velocities of points on both objects. Due to Weber's law, the difference thresholds will be elevated.

The looming detector scheme is made even less tenable by an experiment (Simpson, 1988) in which viewers adapted to a pure Z-rotational flow before viewing a compound (approach + Z-rotation) flow. If looming sensors underlie performance, then adaptation should have no effect since such sensors do not respond at all to rotation. If instead the total flow is sensed in the first stage and then the rotational component is subtracted out to some degree, then we would predict that adaptation should improve discrimination performance over the no adaptation condition. By adapting the sensors which recover the total flow, we reduce their sensitivity to the rotational component. When they are presented with the compound flow, they are less sensitive to the rotational component—it is as if less rotation were present in the flow, and therefore less degradation of performance should result. This is exactly the experimental outcome. Together, the experiments cast doubt on the looming sensor scheme (changing area or line length mechanism) as an algorithm for decomposing retinal flow.

One rather simple way to perform the decomposition of retinal flow would be to use sensors that pick up the rotational flow; what is left over is the translational flow. Such sensors would be easy to construct because the rotational flow does not depend at all on the depths of the objects in the scene. In fact there is physiological and psychophysical evidence for neurons sensitive to the rotational flow.

Neurons sensitive to the components of rotation in optic flow have been found in rabbits and monkeys. Simpson et al. (1979) rotated large (20 × 20 deg) and slowly moving (0.1–1 deg/s) patterns around stationary rabbits. In intracellular recordings
from neurons in the accessory optic system (which projects to the vestibulocerebellum), they found three preferred axes of rotation, each corresponding to a principal axis of a semicircular canal. It seems that the rotational part of the optic flow is decomposed into $X$, $Y$, and $Z$ components. Cells have been found in macaque MST that would be suitable for extracting the rotational component in optic flow (Tanaka et al., 1986). ‘Field-type’ cells respond to wide-field movement (a $40 \times 55$ deg field was used). The stimuli used were approximately $X$- or $Y$-rotations. As for the $Z$-component, Saito et al. (1986) found cells that prefer such rotations.

Probably the most striking psychophysical evidence for channels sensitive to the rotational flow is the phenomenon of vection. Although the translational component in optic flow contains all the depth information, the rotational component can theoretically give the observer information about his self-motion. In fact, a large-field pure rotational flow does give viewers a powerful sensation of self-motion, or vection (Dichgans and Brandt, 1978). This is evidence that the visual system contains mechanisms sensitive to the rotational flow and correctly uses this information to get self-motion.

Regan and Beverley (1985) looked for psychophysical evidence of channels sensitive to $Z$-rotation (curl). They adapted observers to two sorts of random-dot motion. The random-dot field was divided into four quadrants (see Fig. 10). The dots in any quadrant oscillated sinusoidally along a straight line with the same amplitude. The difference between the two types of adapting stimulus was in the relative phasing of the quadrants. In the inphase condition the dots in each quadrant moved diagonally towards the next quadrant. Such a display has a large $Z$-rotational component. In the antiphase condition, although the same vectors were present, the vectors in two opposite quadrants were exchanged. Such a display has no net rotational component. Although both adapting stimuli had the same oscillations present, in one there was no net rotation and in the other there was. The inphase stimulus raised thresholds for detecting true rotary motion much more than did the antiphase stimulus. This led
Regan and Beverley to conclude that channels sensitive to Z-rotation (vorticity, curl) exist.

The same sort of argument that was presented in the discussion of Regan’s looming experiments can be used to explain these results in terms of simple velocity sensors instead of special curl detectors. When the aftereffect of each velocity in the adapt phase is added to the test stimulus, we find that the magnitudes of the resulting perceived motions are smaller when the test and adapt stimuli are both rotating than when the two stimuli are different. This argument is a little less compelling for oscillating local motions because the usual stimulus for motion adaptation is unidirectional. It is not unreasonable to suppose, however, that adaptation to an oscillating motion will elevate motion thresholds for both directions.

If the visual system uses any of the schemes outlined earlier, it can filter out the rotational component of the optic flow when required to recover depth. If it cannot filter out the rotational component, though, we expect impaired ability to discriminate depth as the rotational component increases. Let us consider the case of an X- or Y-rotational component, whose local effect is to add a common vector to all retinal vectors. If two objects are at just discriminably different depths, their retinal vectors are just discriminable. If we now add a rotational component, a common vector is added to both. Now both vectors are longer. Weber’s law predicts that the vectors will no longer be discriminable since the difference threshold rises with the length of the vectors. Nakayama (1981) found just this degradation of speed discrimination performance as a common velocity was added. This experiment reinforces Simpson’s (1988) conclusion that the visual system is unable to filter out the rotational component in optic flow.

This result that adding a rotational component to the flow degrades a viewer’s ability to recover depth is consonant with similar findings on self-motion. Llewellyn (1971), Johnston et al. (1973), Regan and Beverley (1982), and Rieger and Toet (1985) all found that the ability to detect heading from optic flow declines as the size of the rotational component increases. The translational component contains all the information about heading.

The conclusion of this section appears to be that humans can only poorly decompose the flow into rotational and translational components. However it should be pointed out that unless the field of view is very large, this decomposition is in principle very difficult. The reason is that for a small field of view an X- or Y-rotation looks like an X- or Y-translation of a scene with no depth (cf Negahdaripour and Horn, 1989). In these cases the retinal flow will be a set of parallel vectors. It may be that human decomposition of optic flow is poor unless the field of view is nearly 180 deg vertically and horizontally.

3.2.1. The kinetic depth effect. Metzger (1934a, 1934b) was an early worker on what was later dubbed the ‘kinetic depth effect’. A shadow projection of elements undergoing a turntable-centred Y-rotation immediately reveals the 3D layout of the elements. Metzger pointed out that the elements in the projection undergo simple harmonic motion (this is actually only approximately correct for polar projection). That is, the position (or velocity or acceleration) of each dot is a sinusoidal function of time. Varying the amplitude and phase of the dots’ sine functions will change the perceived layout (see Simpson, 1986).
It was Wallach and O'Connell (1953) who named the kinetic depth effect and did some famous experiments on it. White and Mueser (1960) did some similar ones with pegs rotating about a turntable-centred Y-axis. Neither of these studies tells us much beyond the fact that 3D structure can be obtained from optic flow induced by object-centred Y-rotation.

A critical point to consider about kinetic depth effect displays is that even though the stimulus is rotating, all the information about depth comes from the translational component of its optic flow and retinal flow. If this seems contradictory, refer again to Section 2. The coordinate system we are using to describe all motions is centred at the observer's eye. A rotation about some axis other than the eye is not a pure rotation in this coordinate system—it is a rotation and a translation. A pure rotation cannot convey any information about depth—the kinetic depth effect works only because of the translational component, which in turn is due to the object's centre of rotation being displaced from the origin. Figure 11 shows a top view of a random-dot-filled cube in rotation about a self-centred Y-axis. A rotation of the cube by 45 deg about its own Y-axis $(X_c, Z_c)$ is equivalent to a rotation of 45 deg about the origin (middle panel) followed by a translation (bottom panel). In detail, the coordinates of a given point to begin with are $(X, Z)$—the diagram is a top view, deleting the $Y$-coordinate. After rotation of $\theta$ deg around the origin, the new coordinates are (cf Harrington, 1983):

\begin{align}
X' &= X \cos \theta - Z \sin \theta, \\
Z' &= X \sin \theta + Z \cos \theta. 
\end{align}

Then we do the translation:

\begin{align}
X'' &= X' - X_c \cos \theta + Z_c \sin \theta + X_c, \\
Z'' &= Z' - X_c \sin \theta - Z_c \cos \theta + Z_c.
\end{align}

It is the translational component of the object motion that allows the depth of the object to be recovered. In general any motion can be described as a combination of a rotation about the origin and a translation (Chasle's theorem). It is clear from the figure that most of the object's translational component is in the $X$-direction.

In terms of the optic flow, the effect of the $Y$-rotational component is to add the same angular velocity to all points regardless of their depths. It is the translational component that produces retinal velocities that vary with depth. In the retinal flow, the result of a rotation about a self-centred axis looks like the top panel of Fig. 12. This is almost a pure $X$-translational flow. The difference is that the vectors are not quite parallel. This curvature of the vectors' paths is due to the $Y$-rotational and the $Z$-translational components. The relative depths can be obtained from the vectors' lengths.

While we are analyzing the information in kinetic depth effect displays, we should address the issue of parallel and polar projections. A polar projection is the ordinary perspective that occurs in lens-based imaging systems such as the eye. In a parallel projection, the $Z$-coordinate of a point is simply omitted; it does not affect the retinal location of the projected point. Braunstein (1962) and Green (1961) have measured the ability of observers to recover depth in parallel and polar projections. The subjects could get the depth quite readily from the parallel projections.
Figure 11. Top view of a rotation by 45 deg about self-centred Y-axis (0, 4). This rotation of the points (shown in their original positions at the top) is equivalent to a rotation of 45 deg about the Y-axis of the viewer's coordinate system at (0, 0) (middle), followed by a translation of the points to their final positions (bottom).

It is not difficult to explain how a parallel projection can give depth information, though it does involve some rethinking. We have just seen that the depth information in the polar projection of the display comes from the translational component; the rotational component conveys no depth. For a parallel projection, though, the situation is reversed: it is the rotational component rather than the translational component that gives depth. In a parallel projection, since the Z-coordinate is omitted, the retinal velocity of a translating point does not depend on its distance. The retinal velocities resulting from a translation will be the same for all objects (in fact, only the part of the translation parallel to the projection plane will be visible). Clearly a translation can give no information about depth if the projection is parallel. On the other hand, the retinal velocities resulting from parallel projection of a Y-rotation do depend on the depths of the points. In a parallel projection the projected coordinate x is identical to the world coordinate X. Examination of Eqn (26) reveals that two
points with the same $X$ coordinate but different $Z$ coordinates will acquire different $X$ coordinates after a given rotation. Their $X$-velocities (and $x$-velocities) are therefore different; the retinal velocities resulting from the parallel projection of a $Y$-rotation depend on the depths of the points.

So does the brain then use one set of computations for parallel projections and another set for polar projections? That seems rather unlikely. Instead the brain would interpret the parallel projection as having arisen from a polar projection, since that is what it has seen over its whole lifetime. Compare the retinal flows arising from a polar projection and a parallel projection of the same rotation about an object-centred $Y$-axis (Fig. 12). The difference between the two is that the parallel projected flow consists of parallel retinal vectors instead of vectors lying on curved field lines. The lengths are the same. In fact, the bottom panel of Fig. 12 shows a pure $X$-translational flow. In such a flow the relative depths are given straightforwardly by the lengths of the vectors. And this is how the visual system interprets the flow.

Having discussed the depth information available in flows from self-centred $Y$-rotation, let us now review some of the psychophysical results. Many studies of the kinetic depth effect have used dependent measures that are seriously flawed. Green (1961) and Braunstein (1962) both asked subjects whether the stimuli looked 2D or
3D, and asked for coherence judgements. The 2D vs 3D judgements are hard to interpret given Braunstein’s (1966) finding that half of his subjects viewing objectively 2D and 3D textures undergoing a pure Z-rotation (a transformation yielding no depth information) called the displays 3D! The coherence rating is also hard to interpret because there was no manipulation of the physical coherence of the display. In motion work coherence typically refers to the correlation of elements across time (cf van Doorn et al., 1985); coherence is reduced by adding noise.

The objective coherence or correlation of moving stimuli was varied by Lappin et al. (1980) and Doner et al. (1984). In both cases, noise dots were added to sequences of a sphere rotating about its own Y-axis. In the first study it was shown that the ability to discriminate correlation (i.e. to tell which of two sequences had more added noise) in two-frame sequences rapidly declined as the correlation of the standard fell below 1.0. The study by Doner et al. showed that the depth-from-motion system’s susceptibility to noise decreased with the number of frames. This implies an integration of information across frames. It is not clear that this experiment tells us something special about the noise resistance of the depth-from-motion process as opposed to the noise resistance of the initial registration of 2D velocities. Threshold signal-to-noise ratios for 2D motion have been extensively measured by Koenderink and associates (for a review see van Doorn et al., 1985).

Siegel and Anderson (1988) presented Rhesus monkeys and humans with a parallel projection of points on the surface of a cylinder that was rotating about its own Y-axis. The subjects’ task was to release a lever when the moving sequence changed from a control sequence having no 3D structure to the 3D sequence. Psychometric functions were obtained for ‘fraction structure’ (something not clearly explained in the paper), point life (each individual point ‘dies’ after a specified number of frames and is replaced at a new random position), and number of points. The functions for the human and monkey subjects for whom data is given are similar. Note, however, that it is not always the same individuals whose data are being compared and also that not all subjects’ data are shown (only one human and one monkey psychometric function are shown in each plot).

Sperling and colleagues have recently pointed out flaws in previous work on the kinetic depth effect, and have promoted a new method of measuring depth perception in such displays. Sperling et al.’s (1989) criticisms amount to this: responses by subjects in previous studies were interpreted by the experimenters as measures of depth perception, but these responses could have been based on simple 2D cues in the displays. Sperling’s displays attempt to eliminate these cues. On each trial the subject is presented with one of 53 dot-covered shapes that is rotating about its own axis. The task is to identify the shape.

It is hard to summarize existing psychophysical work on depth recovery from flows produced by object-centred Y-rotation. The basic conclusion is that depth can be recovered from such flows.

**Algorithms for depth from object-centred rotation.** There are schemes that can only analyze flows arising from object-centred rotations (kinetic depth effect displays). The most well-known of these is due to Ullman (1979a, 1979b). Ullman’s structure-from-motion theorem takes as input the point correspondences from a parallel projection of an object rotating about its own axis. Other such schemes have been proposed by
Webb and Aggarwal (1981) and Hoffman and Bennett (1986). There may be times when it is helpful to use a special-purpose algorithm such as these. However, the schemes discussed in previous sections can deal with a completely general motion, and it seems more reasonable to use a method for extracting depth from motion that can be applied to all possible flows.

3.2.2. The stereokinetic effect. The dadaist artist Marcel Duchamp invented a form of kinetic art in which a 2D pattern (usually composed of concentric circles with gradually shifting centres) when rotated about an axis parallel to the viewer's Z-axis yields a strong depth impression (akin to a stereoscopic view). Musatti (1924, 1931) labelled this the 'stereokinetic effect'. Subsequent studies include Wallach et al. (1956), Fischer (1956), Wilson et al. (1983, 1986), and Robinson et al. (1985).

How can Z-rotating a flat pattern give a 3D impression? Wallach et al. and Robinson et al. have previously given formulations similar to the one I am now proposing, although I will use the optic flow terminology introduced earlier. A pure rotation about a viewer-centred axis gives no depth information. However, in the stereokinetic effect, a pure Z-rotation of a 2D pattern gives rise to the perception of objects moving in depth. This conflict can be resolved if we assume that the pure rotation is not sensed as a pure rotation but as a rotation and a translation (the latter giving the depth). As always, by 'translational component' in the retinal flow I do not mean a simple translation of all points in the image plane. I mean the retinal flow resulting from a translation relative to the scene.

The starting point for our explanation, then, is that the motion sensors do not yield a retinal velocity field corresponding to the optic flow stimulus. The derivation of such a nonveridical velocity field by the human visual system, as for example in the barbershop pole illusion, has been noted before and is treated thoroughly by Hildreth (1984). The nonveridical motion is due to the 'aperture problem': if the motion sensor's receptive field (aperture) is small compared to the extent of the moving edge it is looking at, then it can only detect the component of the motion perpendicular to the orientation of the edge. Think of a textureless circle spinning about its own centre—clearly no motion can be detected since there is no motion perpendicular to the contour. If the circle rotates around an off-centre axis, the normal component of the edge's motion can be detected. Hildreth's algorithm for computing the velocities of edges produces velocity fields that are different from the true velocity fields (a translational component is added) and consistent with human perception.

Let us then explain the stereokinetic phenomenon using the aperture problem. The display shown in the left of Fig. 13 is Z-rotated about the centre of the largest circle. Because of the aperture problem, the retinal velocities of points on the largest circle will be nil. As we go from the outside to the inside, the centres of the circles gradually shift. Thus the size of the normal flow gradually increases as we go from the outside to the inside. This normal flow is not identical to the true pure Z-rotational flow. There is a Z-rotational component, but there is also an XY-translational component. The size of the retinal translation vectors increase from the outside to the inside of the display; their lengths give the depth. The longer the retinal translation vector, the closer that part of the display will seem.

Figure 13 shows the stereokinetic display at two successive instants where a large X-translational component is created by the aperture problem. By viewing the figure
Figure 13. Two successive views of a stereokinetic pattern Z-rotating about the centre of the largest circle. Because of the aperture problem, the velocity sensors can only respond to the component of the true flow that is normal to the edges of the circles. This normal flow includes a substantial translational component. (Note: This does not mean uniform translation in the image plane.) The size of the normal flow increases towards the centre of the display as the centres of the circles become increasingly displaced from the centre of rotation. Thus the retinal translation of the circles increases towards the centre. The lengths of the translational vectors give the depth. Cross-eyed fusion of the display will reveal the depths seen during motion of the display. Binocular stereopsis uses the translational component (disparities) to compute depth; the aperture problem also affects stereopsis.

cross-eyed, the reader can confirm that the retinal translations (known as disparities in the binocular literature) increase from the outside to the inside of the display. The centre circle has the largest retinal $X$-translation and appears closest. In general the translation will have both $X$- and $Y$-components. The successive instants were chosen to maximize the $X$-component for the benefit of the binocular stereoscopic system, which cannot handle $Y$-translation. I must stress again that the successive views in Fig. 13 were generated by rotation in the plane about the centre of the pattern. The translation in the retinal flow is caused by the aperture problem. This problem would be as prevalent in binocular stereo as in motion, but as far as I know it is completely unexplored there.

3.3. Multiple interpretations of optic flow

Illusions, or more accurately ambiguities, in human derivation of depth and motion from optic flow have been reported several times. There are two main classes of these ambiguities. The first class has to do with the depth seen in parallel projections of objects rotating about their own axis. The explanation for the ambiguity here is trivial. The second class of ambiguity has to do with the depth seen in planar displays. These are rather more difficult to explain.

As noted by Sinsteden (1860), Kenyon (1898), and Miles (1931), the blades of a fan (or windmill) rotating in depth and seen in silhouette seem spontaneously to change direction of rotation. Under the conditions used, the projection to the eye was essentially parallel, so it is not surprising that the direction of rotation about the object-centred axis was ambiguously perceived. The angular velocity of a ray connecting a point on the fan and the eye is not appreciably different when the point is on the fore or aft part of its trajectory (cf Cutting, 1986, p. 191). Thus the depth information is ambiguous; it is given that the point is changing depth, but not the
direction of change. As Ullman (1979a) puts it, a parallel projection of rotation about a self-centred axis can give the depth only up to a reflection (i.e. a reflection in depth).

The second class of depth ambiguity is described by Gibson et al. (1959). They projected the shadows from two transparent sheets sprinkled with talcum powder onto a viewing screen. The two sheets were set a variable distance apart and moved rigidly together sideways (parallel to the viewer’s X-axis), casting points moving at two velocities. At the largest separations the viewers reported seeing two planes in depth, but the depth assignment—which was seen in front—was variable. However, when one slanted plane (yielding a gradient of screen velocities) was moved the same way, all viewers saw the correct gradient of depths. Gibson et al. concluded that it was necessary to distinguish between motion parallax (the two-velocity case) and motion perspective (the many-velocity case). The former is the classical depth cue, and not sufficient for depth localization; the latter is the more ecological situation, and is sufficient.

The results of this experiment cannot be explained away, as Prazdny (1981b) tries to do, by saying that the depth cannot be found for the two translating planes because the surfaces are moving independently, violating the rigidity assumption. This is simply not true; the surfaces were moving rigidly together. One trivial explanation might be that it was difficult to say which plane was in front, but it was easy to see which was in front. The two planes were similarly textured, and the authors do not say how the viewers were to indicate which was in front. Viewers could not say, for example, “the random textured one is in front” since both sheets were randomly textured, or “the red one is in front” since the dots on both sheets would cast black shadows. In fact the two sheets seem to have been alike in every way, so it would have been very difficult to verbalize which appeared in front.

Another possibility is that the visual system sometimes uses a method for getting depth from motion that is good only up to a reflection (Koenderink, 1986, outlines such a method). By this I mean that the depth recovery mechanism does not distinguish reflections in depth—e.g., a plane inclined 45 deg towards the viewer and one inclined 45 deg away, or a convexity and a concavity. Helmholtz (1925, p. 287) notes that a relief is often hard to distinguish from the same figure in the round. A video by Proffitt et al. (1987) contains a nice demonstration of this illusion using a moving view of the inside (relief) of a mask. Although the optic flow should ideally reveal the true relief of the mask, we incorrectly see it as pointed outwards (like a sculpture in the round) instead of in relief (pointed inwards). Note, however, that the mask is almost completely textureless and thus the motion will be undefined over large areas of the retinal image.

There are other differences between the two-transparent-planes display and the sloping-plane display besides the range of velocities present (2 versus a large number). The most important difference is that the flow arising from the single sloped plane is single-valued, whereas in the two-transparent-planes display, there may be two distinct retinal vectors at a given point. This may cause serious difficulties. A better comparison would have been the sloped plane versus a step. However, if this comparison is done, one will again see Gibson et al.’s result: the plane is vividly seen to slope in depth, but the step looks two-dimensional. It may simply be, as Gibson et al. suggest, that more than two depth planes are necessary to give unambiguous depth
percepts. Displays with three depth planes give a vivid depth impression, with no depth reversals or ambiguity.

It turns out that the flow from a plane is particularly difficult to interpret on the theoretical level. Longuet-Higgins (1984) treats the topic thoroughly. In addition to the correct interpretation, a flow from a plane has three spurious interpretations. But this would not be the explanation of the Gibson et al. results, because the spurious interpretations occur only for an instantaneous flow. After extended viewing, the spurious interpretations can be ruled out. Note, however, that planar surfaces pose difficulties for some algorithms (e.g., the one proposed in Section 3 of Longuet-Higgins and Prazdny, 1980).

Depth-from-motion ambiguities also occur for flows from curved surfaces. Waxman and Wohn (1988) found that ovoid surface patches almost always have a unique 3D interpretation, but cylindrical and saddle surfaces can have two or three spurious solutions respectively.

3.4. Nonrigid optic flow and motion segmentation

This paper has considered only optic flow arising from rigid motion. Although many flows in everyday life are generated by rigid motion, many are not. Nonrigidity can occur in two major ways. First, an object can deform while being viewed. The deforming object can be piecewise rigid (hinged or jointed) or it can bend or stretch (e.g., a jellyfish’s body while it swims). The second type of nonrigidity occurs when we move relative to a world that is mostly stationary but has some independently moving objects.

Johansson (1975) has found that the shape of a moving human being can be recovered from a view of several light sources attached to the joints. Many others have studied the perception of such ‘biological motion’. Computer vision approaches have treated this stimulus as a special case. For example, Hoffman and Flinchbaugh (1982) discussed motion of linked hinged rigid rods (limbs) constrained to move in one plane. Obviously an approach like this would not recover the 3D structure of a pulsating jellyfish.

Depth can also be recovered from the even more nonrigid flow resulting from bending or stretching a rubbery sheet (Jansson and Johansson, 1973; Jansson and Runeson, 1977). Computational schemes for dealing with such flows exist. Ullman’s (1984) incremental rigidity algorithm works best for rubbery objects that are almost rigid (though it can be applied to any motion). Shulman and Aloimonos (1988) develop a related idea. Koenderink (1986) presented an algorithm that can recover shape in the presence of bending deformations, and his informal observations suggest that the algorithm is consistent with human performance.

The problem of extracting depth from motion relative to a stationary scene is complicated by the presence of independently moving objects. This sort of scene is piecewise rigid. One approach is to segment the retinal flow into several regions, each containing a view of a single rigid object. The depth from each region of rigid flow can then be obtained. This general approach is used by Adiv (1985), Murray and Williams (1986), Maybank (1987, Section 6) and Murray and Buxton (1987).

There has been no psychophysical work on 3D displays of independently moving objects. It is well known, however, that motion can be used to segment the retinal image. For example, viewers of a random-dot kinematogram can readily identify the
shape of the region of coherently moving dots (Braddick, 1974; Baker and Braddick, 1982). It seems plausible that observers can segment regions by the retinal velocities resulting from 3D motion.

### 3.5. How is depth from optic flow represented?

Marr (1982) distinguished between the 2D sketch and a full 3D model representation. The 2D sketch gives the depth, orientation, and discontinuities of depth or orientation of the surface in a viewer-centred reference frame. The full 3D representation is an object-centred description of the object shape. The research reviewed here on the representation of depth from optic flow is mostly at the level of the 2D sketch.

The simplest possible representation would be of the raw relative depth values. A slightly more sophisticated representation would be in terms of surface orientation or curvature (both of which, we have seen, are potentially available from the retinal flow).

One possible basis for representing depth from motion is spatial frequency. The idea is essentially the same as that of spatial-frequency analysis of spatial luminance distributions (cf. Cornsweet, 1970). To get this representation, a 2D Fourier transform of the raw depths in each visual direction is done. Other possibilities besides a global Fourier representation would include representations derived by windowed Fourier transforms, wavelet transforms, or Gabor transforms.

The idea of a depth representation based on frequency seems implicit in Rogers and Graham’s (1982) investigation of depth from motion and from binocular disparity. The sensitivity of the visual system to sinusoidal depth gratings (conveyed by motion or disparity) at various spatial frequencies was measured. The shapes of the obtained sensitivity functions were identical for both disparity and motion gratings, but in two of the three observers the thresholds for motion gratings were a factor of two higher (the results were plotted in equivalent units). The motion contrast sensitivity function was bandpass, which would be consistent with motion cells having Mexican-hat receptive field profiles (e.g., excitatory for a given direction of motion in the centre, inhibitory for that direction of motion in the surround). Such cells were postulated by Nakayama and Loomis (1974) and have been found in monkey MT cortex (Allman et al., 1984).

All of the possible representations discussed so far—raw depth, slope, curvature, spatial frequency—assume that the depth recovered has a metric property. Todd and Reichel (1989) argued from their depth-from-shading experiments that the depth representation is not metric but only ordinal. That is, depths of parts of the scene are rank-ordered. Todd and Bressan (1990) did motion experiments using a 2-frame parallel projection of shapes rotating about a self-centred Y-axis. They found that people cannot discriminate two objects if one is a Z-stretched version of the other. Since this is an affine stretching transformation, Todd and Bressan concluded that the human representation of depth obtained from motion may be based on affine structure.

One set of results at odds with this conclusion comes from Lappin and Fuqua (1983). Their subjects’ task was to judge whether a point bisected the distance between two other points in 3D space; the 3D configuration of points was in motion when viewed. Clearly the task requires 3D metric distance information, and the subjects
performed quite well. Lappin (1990) argues that there is an object-centred metric representation (which is not the same as relative depth).

3.6. Integrating depth information from other sources

Marr's legacy in computer vision has been a large number of papers on obtaining depth from various sources: shading, binocular disparity, contour, texture, and motion. Ultimately, we will want to combine the depth information obtained from all sources. Marr (1982) proposed that the outputs (distance from viewer, depth and orientation discontinuities) of the various depth modules were combined to form an object-centred 3D representation. The approaches to be treated here are not concerned with such a high level of representation. Instead they just try to combine depth from different sources; constructing a 3D model is a problem that can be postponed.

One approach to the problem of combining depth information is regularization, which assumes that the physical processes giving rise to the image, including the depths of scene points, change slowly over space (Poggio et al., 1985). Essentially, the information for depth coming from each cue is accumulated. One problem with the regularization approach is that surfaces are smooth in most places, but not at the edges; the same is true for their images. Edges are highly informative aspects of an image. For example, it is likely that discontinuities in the image colour, luminance, texture, disparity, and motion will coincide at an object boundary. Poggio et al. (1988) used Markov random fields to integrate information about surface discontinuities.

Aloimonos (1989) and Aloimonos et al. (1988) presented an interesting method for recovering the orientation of a planar surface by combining information from shading, texture, and motion. Aloimonos pointed out that the shape from shading, contour, texture, or motion problem is simplified if the observer is active. Some behavioural evidence supports the idea that information about the observer's self-motion is used to derive depth. Goodale et al. (1990) trained gerbils to jump over a gap of variable size. Before jumping, the gerbils generated retinal motion by bobbing their heads vertically. Goodale et al., argued that the depth estimate was derived by combining information about the velocity of the head bob with information about the velocity of the image motion. Ono and Steinbach (1990) found that subjects report seeing more depth in X-translational displays when the display motion is yoked to the observer's head movement than when there was no such yoking.

The question of depth integration has been given increasing attention in psychophysics. Bülthoff and Mallot (1990), for example, did experiments on depth integration of disparity, shading and texture. Maloney and Landy (1989) proposed that a simple linear combination of depth estimates from the various cues takes place. This proposal is supported by the results of Dosher et al. (1986), who found that depth from linear perspective, binocular disparity, and 'proximity luminance covariance' (brighter areas appear closer) combined additively when subjects viewed Y-rotating Necker cubes. Landy et al. (1991) also found that depth perception takes a weighted average of information conveyed by texture and motion.

Optic flow can only give relative depth. An important question is how information for absolute depth such as given by accommodation or convergence is combined with that from optic flow. Such calibration is necessary for using optic flow to regulate action (although as Lee has pointed out, TTC could be used as such absolute information). Ono et al. (1986) presented motion parallax displays; the subjects sat at
various distances from the monitor (and thus had access to information from convergence and accommodation). As we would expect if depth from motion were scaled by absolute distance information, the perceived depth in the display increased with viewing distance.

Regan has found evidence for the integration of two sorts of dynamic depth information: changing size and changing disparity. We have already discussed the evidence for changing size mechanisms (or looming sensors). A changing disparity mechanism responds to binocular flows—flows that are the difference of the retinal flows sensed by each eye (see Waxman and Duncan, 1986, and Waxman and Wohn, 1988, for an analysis of the information available). Such a disparity flow gives the rate of change of disparity at each point. Beverley and Regan (1973, 1975) have found psychophysical evidence for the existence of changing disparity channels, and Regan et al. (1979) found such neural units in the cat. Regan and Beverley (1979) showed that a changing disparity aftereffect could be cancelled using changing size and vice versa. This demonstrates that both changing size and changing disparity feed the same motion-in-depth mechanism.

4. CONCLUSIONS
Here is a brief summary of some of the major points covered:

- the information for depth is contained solely in the translational component of optic (or retinal) flow. To get depth, the rotational component must somehow be filtered out. But the experimental evidence indicates that humans do this rather poorly.
- surface slant, tilt, and curvature may be recovered through the use of first and second spatial derivatives of retinal velocities. This may prove to be impractical with real images, and human experimental results make it unlikely that such spatial derivatives are computed in the human visual system.
- retinal velocities are not the only possible input to the depth-from-motion process (other candidates include successive snapshots or point correspondences, changes in texture density, line length, and area). However, retinal velocities are the most plausible input given the experimental evidence.
- the kinetic depth effect and the stereokinetic effect convey depth through the translational component in the optic flow arising from them.
- the human visual system appears to be most comfortable in extracting depth from the most ill-behaved retinal flows. Better behaved flows (e.g., arising from planes and simple curved surfaces) can have multiple interpretations. The multiple interpretations have been proved theoretically and observed experimentally.
- the problem of segmenting flows arising from multiple moving objects is almost untreated in the psychophysical literature. A similarly ripe field (though it has received some attention) is the nature of the 3D representation derived from motion.
- depth from motion seems to be combined in a quite straightforward way with depth obtained from other sources.

There has been great recent progress in understanding depth from optic flow. Yet we are still a long way from being able to build a practical computer vision system for deriving depth from motion, or from understanding precisely how the human brain manages to do it.
ACKNOWLEDGEMENTS

This work was supported by operating grant 46579 from the Natural Science and Engineering Research Council.

REFERENCES


Bourdon, B. (1898). La perception monoculaire de la profondeur Révue Philos. 46, 124–145.


Regan, D. and Beverley, K. I. (1982). How do we avoid confounding the direction we are looking and the direction we are moving? *Science* 215, 194–196.


