Sperling, G., and Dosher, B., A. (1994). Depth from motion. In *Early Vision and Beyond.* Cambridge, MA: MIT Press, 1994. Pp. 133-142.

Depth from Motion

George Sperling and
Barbara Anne Dosher

Overview

This chapter is concerned with the perception of object depth and object structure that results from monocular viewing as distinct from stereoptic depth that results from binocular viewing. Two successive views of a rotating object contain precisely the same kind of information as the two views from different eyes. As an object rotates, there is a continuum of views. Therefore, there is no physical reason why the depth that is perceived in viewing a rotating object might not appear to be even more realistic and more depthful than stereoptic depth.

We will demonstrate here that it is the motion flow field that contains the information that is used to extract perceived depth from dynamic monocular displays. Once a display is perceived as having three-dimensional (3D) depth, its apparent rigidity—or lack thereof—is a property derived from the successive 3D structures over time, rather than vice versa, as would be suggested by algorithms that use rigidity itself to extract the depth structure (e.g., Gryzwacz and Hildreth, 1987; Grywacz et al., 1988; Hildreth et al., 1990; Bennett et al., 1989; Koenderink and van Doorn, 1986; Longuet-Higgins and Prazdny, 1980; Ullman, 1984). In other words, we hypothesize that the sequence of perceptual computations is first *depth from motion,* then *structure from depth.*

Motion-Depth-Sign Ambiguity

The computation of depth from object motion is inherently ambiguous with respect to the sign of the depth (the *motion-depth-sign*). Without an additional cue (such as knowledge of the self-motion that produced the motion flow field), a particular motion flow field supports two plausible depth isomers, with depth relations that are mirror reflections of each other (e.g., Ullman, 1979). A first step in computing depth from motion is the choice of one of these two possible depth isomers. For example, when self-motion produces the object motion (motion parallax), the perceptual motion–depth computation is immediately disambiguated (Ono and Steinbach, 1990; Ono, Rivest, and Ono, 1986). Many cues can exert a powerful influ-

ence on the motion-depth-sign. The decision mechanism for determining the motion-depth-sign is like a balance scale that tilts in either one of two directions. We will show that cues favoring one depth isomer versus another exert their influence by combining additively, just like weights in the two pans of the balance scale.

The potency of a depth-disambiguation cue is greatest during the very first instant of viewing a dynamic display. Once a particular motion-depth isomer is perceived, it is quite stable and resistant to change. This path dependence is characteristic of winner-take-all cooperation–competition neural networks (e.g., Sperling, 1970) and also of systems of dipoles such as those responsible for the magnetic properties of metals (Julesz, 1971). Some of the consequences of this kind of computation are considered below.

Size Indeterminacy of Monocular Vision

Because the distance between the eyes is known, stereopsis can give absolute depth information; for example, we can use stereopsis to thread a needle. Monocular vision without head or body movements, and neglecting accommodation, is geometrically excluded from yielding absolute size information. A firefly in front of the eye and a cataclysmic solar event could cast identical retinal images.

Depth-Scale Indeterminacy of Instantaneous Flow Fields

Indeed, the *instantaneous* motion flow field fails not only to yield absolute size, but also to yield the depth scale of a shape: An instantaneous flow field could have been generated by a very small movement of a very depthful object, or a larger movement of a relatively flatter object (cf. Adelson, 1985). However, under ordinary circumstances, two different flow fields suffice to yield the depth scale. That is, once the angle of object rotation around an axis perpendicular to the line of sight is greater than zero (in a noiseless system), depth-scale ambiguity is optically resolvable. When the angle of rotation reaches 90°, the object that was initially viewed end-on has rotated to appear sideways, and the 90° image provides perfect shape resolution. The minimum angle of rotation required for perceptual resolution of depth-scale ambiguity depends, of course, on the shape of the object and the quality of the image.

Processing Architecture: Common Depth Channel

The perception of depth from 2D projections without stereo, shading, or parallax is the *kinetic depth effect* or KDE (Wallach and O'Connell, 1953). When one views KDE stimuli such as rotating Necker cubes, especially in a setting that removes incidental cues to the depth of the display screen on which they are represented, the sense of depth organization for the perceived object is quite vivid —as vivid as the depth organization supported by stereopsis. That perceived depth from kinetic cues can be as realistic as perceived depth from stereopsis suggests that both sources of information feed into a *common depth channel.*

A proposed architecture for the relations and interactions between the various cues to depth and shape is indicated in figure 13.1. The top channel of figure 13.1 indicates a motion signal (a motion flow field) being processed to extract depth from motion. The sign (+ or −) of the extracted depth is indeterminate and this indeterminacy is resolved by a *bistable motion–depth inverter* that multiplies the extracted depth-from-motion relations by either +1 or −1.

Depth from stereopsis is computed in parallel with depth from motion. The depth values computed by stereopsis influence the bistable motion-depth inverter, and, together with other influences, determine the sign (−1 or +1) of the inverter.

Luminance-contrast also influences the bistable motion-depth inverter: High-contrast objects and regions tend to be perceived as being closer than low-contrast objects. The influences from stereo and from luminance add linearly to produce a net influence (e.g., Dosher, Sperling, and Wurst, 1986). Together, motion and stereo depth signals determine a *joint depth map.* This is an assignment of a depth value to each point in the cyclopean x, y image. Just how the depth information from motion and from stereo jointly determine the joint depth map is an interesting, not fully resolved problem.

Depth information also is potentially available from texture, from shading, and from other kinds of inputs. Whether these sources of depth information influence the bistable motion–depth inverter is not yet known. On the other hand, they apparently do influence the joint depth map—at least, that is one way to interpret the findings of Maloney and Landy (1989), who studied the apparent shape of surfaces defined by various combinations of these cues.

A depth map is merely an input to higher perceptual processes, it does not produce any output directly to effectors. The particular higher perceptual process of concern here is object perception, represented in figure 13.1 as an object module that derives object structure from its various inputs. Information encoding at the level of object module is not in terms of an x, y, z, t depth map but in

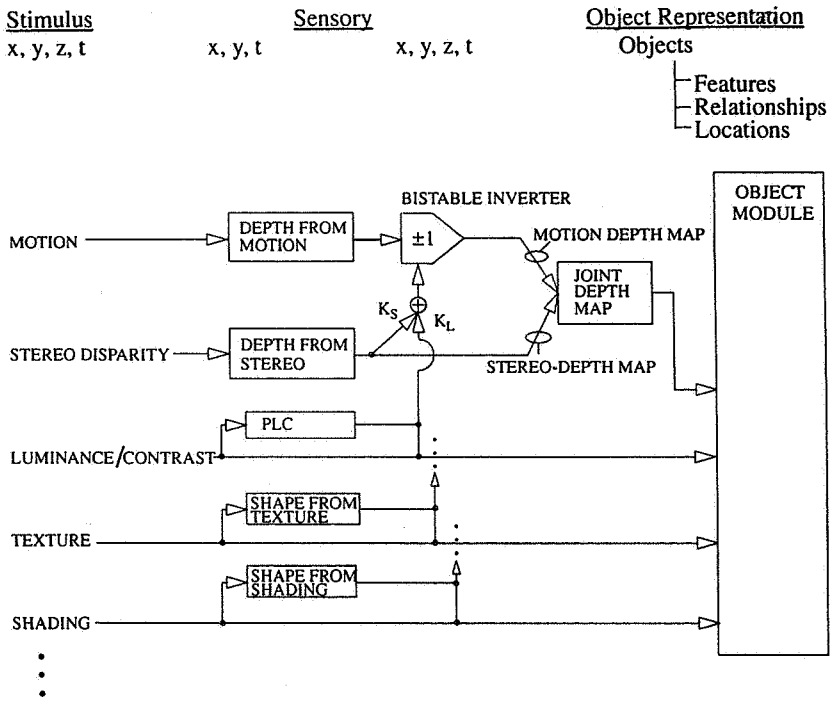

Figure 13.1
Flow chart for the computations of depth recovery from sensory cues and for its relation to the representation of features in an object-memory. External visual stimuli (neglecting color) are represented by their luminance as a function of space x, y, z, (z is the depth dimension) and time t. On the retina, a stimulus is merely a function of x, y, t. For a stationary observer, estimates of the lost dimension z are produced by the depth-from-motion and depth-from-stereo computations. The depth ambiguity of the depth-from-motion computation is resolved by the bistable depth inverter module, which receives additive inputs from stereo disparity, from luminance-contrast computations (proximity luminance covariance, PLC), and perhaps from other cues to produce a joint depth map. Objects are represented as lists of features, relationships, and location; cues (such as motion) are represented here directly in the object module as an object feature, in addition to any depth values that may have been computed from it. It is hypothesized that object rigidity is first computed at this stage.

terms of object components (e.g., geons or volumetric units, Biederman, 1987; Pentland, 1989; Pentland and Sclaroff, 1991) in which depth relations are represented as the shapes of object components (features) rather than as x, y properties of the object as a whole.

Undoubtedly, there is feedback from the object module back to the computations that provide the inputs. For example, the perceived depth isomer of ambiguous objects tends to alternate (Cornwell, 1976; Orbach, Ehrlich, and Heath, 1963; Spitz and Lipman, 1962) and, in our architecture, this would most naturally be implemented by feedback from the object module to the bistable motion–depth inverter.

Finally, it should be noted that in addition to their contribution to a depth computation, motion, texture, shading, and other inputs are represented directly as attributes or features of objects.

We now consider some of the properties and psychophysics of perceptual shape recovery from dynamic visual displays—evidence for the assertions made above.

The Linear Addition of Cue Strengths

The Multistable Perceptions of Necker Cubes

The 3D object interpretation resulting from a 2D projection of 3D rigid-object motion is generally ambiguous, corresponding to two stable perceptual depth isomers. We illustrate this with a Necker cube, a wire object that can be perceived as either one of two depth isomers. The probability of seeing a simple wire Necker cube as one versus the other isomer reflects biases and incidental cues. Among the cues that can disambiguate depth ordering, stereopsis is one of the strongest. Self-motion (motion parallax) also provides a powerful disambiguating cue. A third cue that is sufficient to affect the determination of a depth isomer is luminance contrast: regions of high contrast tend to be perceived as in front of low-contrast regions.

The above cues to depth work about equally well to determine the perceptual depth isomer of a Necker cube

whether it is displayed with orthographic (parallel) projection or perspective projection. In orthographic projection, the two depth isomers are mirror equivalent, so there is no intrinsic reason to choose one over the other. However, when a Necker cube is displayed in perspective transformation (or when a real Necker cube subtends a greater-than-zero visual angle), the two depth isomers are not the same shape: one is a rigidly rotating cube and the other is a nonrigid truncated pyramid, the degree of nonrigidity being proportional to the degree of perspective. The remarkable fact about the perceptual bistability of such rotating Necker cubes is that rigidity per se has little effect on the probability of perceiving one depth isomer versus another. For example, in experiments of Schwartz and Sperling (1983) that used rotating Necker cubes viewed in perspective projection, the probabilities of perceiving the rigid and nonrigid depth isomers were about equal (in uniform-contrast displays). While rigidity had a too-small-to-measure influence on the sign of motion-depth, a minor cue, such as contrast, was highly potent. In these rotating Necker cube displays, differential contrast induced the high-contrast-forward depth isomer in 95% of presentations, for all subjects, whether it was the rigid depth isomer or not.

Jointly Independent Cues to Depth

An objective method for comparing the strengths of the various cues to depth in KDE displays was developed by Dosher, Sperling, and Wurst (1986), who first proposed that multiple cues to depth combine linearly. These investigators studied kinetic depth involving the collateral cues of stereopsis and what they called "proximity-luminance covariation" (see figure 13.2). They presented perspective Necker cubes rotating around a vertical axis to subjects who were asked to report on the direction of rotation (front to the right or front to the left). In such displays, the direction of rotation is apparent immediately at the onset of rotation, and it is perfectly coupled with the either the rigid or the nonrigid perceptual mode.

The two collateral cues to depth organization are schematically illustrated in figure 13.2. The contrast manipulation (proximity luminance covariation) changed the intensity $I(j)$ of the line j in the Necker cube according to several different front-to-rear luminance-falloff schemes. When contrast falls off from front to rear (figure 13.2b), it favors perceiving a rigid cube; when contrast falls off from rear to front (figure 13.2c), it favors perceiving a truncated pyramid. The second cue was rotational stereo, illustrated in figure 13.2e. Stereo disparity could either agree with the generating Necker cube (the rigid isomer) or with the nonrigid truncated pyramid.

A Thurstone Case 5 Model for Predicting Perceptual Mode

Dosher et al. (1986) found that the probability of perceiving the rigid cube isomer was jointly determined by the two cues, and could be accounted for by an additive model that is equivalent to Thurstone's (1947) Case 5 (figure 13.2g–k). Let p be the probability of perceiving the rigid depth isomer. In the absence of any cue, p is assumed to be determined by an individual bias and by internal noise. Internal noise is the standard against which all cues are measured and it is assumed to be normally distributed around zero with unit variance. Positive values correspond to the rigid perception and negative values to the nonrigid perception. Individual bias corresponds to a shift of the noise distribution, corresponding to a tendency to perceive either the rotating cube as rigid or nonrigid in the absence of any other cue. The strengths of the stereo cue and the contrast cue are then estimated in isolation for seven different levels of each, in terms of how far they shift the noise distribution (figure 13.2j). According to the model, when the stereo and contrast cues are now combined in the 7×7 different combinations, their strengths simply add, whether the cues be in concert or in opposition. In fact, this additive strength model provided essentially perfect predictions of the probability of perceiving the rigid cube versus the truncated pyramid for the 7×7 cue combinations.

Why Does an Additive Cue-Strength Model Work so Well?

One way to conceptualize the two stable perceptual states (rigid, nonrigid) for a perspective Necker cube is in terms of an energy map (Sperling, 1970; Sperling and Dosher, 1987). The energy map in figure 13.2k represents the perceived nonrigidity under rotation of all the perceptual 3D reconstructions—rigid and nonrigid—of the Necker cube as indexed by their perceived front-to-rear depth. The minima (energy wells) represent the two rigid depth isomers under parallel projection. Assume that at the instant a display depicting a rotating cue is turned on, the perceived front-to-rear distance is zero. This perceptually flat Necker cube would appear to be highly nonrigid under rotation. The high degree of nonrigidity is represented by the high ridge between the two energy wells. A marble represents the current state of the system; at the onset, it is delicately perched on the ridge dividing the two energy wells. At this instant, the system is in highly unstable equilibrium. Cues are represented as forces acting to push the marble in one direction or the other. Forces combine linearly, but once the marble is in

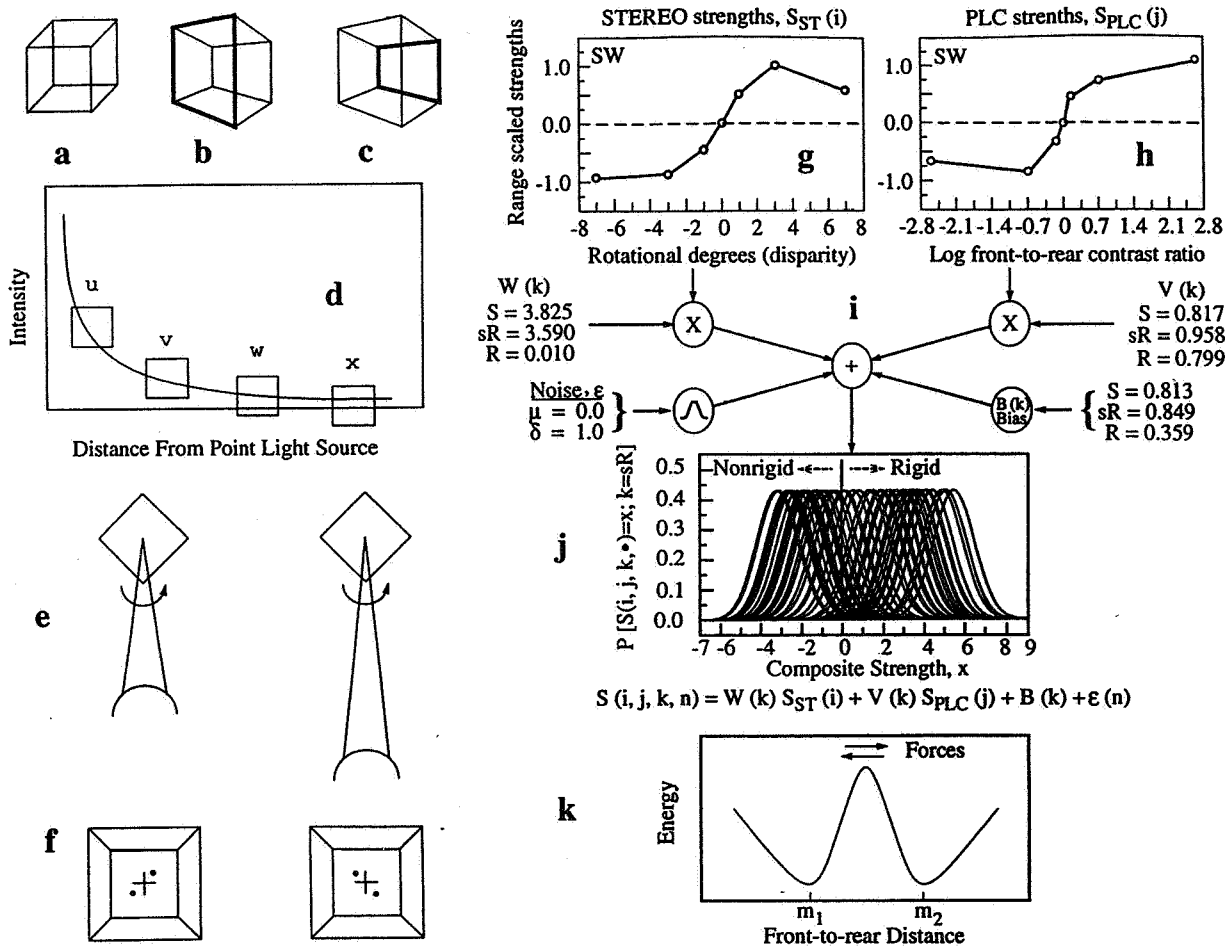

Figure 13.2

The perceived depth isomer of a Necker cube is predicted by a model of additive cue integration. (*a–c*) Necker cubes: (*a*) has neither perspective not contrast cues; (*b*) contrast agrees with perspective; (*c*) contrast opposes perspective. When (*c*) rotates in 3D, the dominant percept is a nonrigid truncated pyramid. (*d*) A diagram to indicate how intensity (or contrast) would fall off from the front to the rear of a self-luminous Necker cube at various distances. Within (*d*), the strength of contrast cue is $(u) > (v) > (w) > (x)$. (*e*) illustrates two simulated extents of stereo; large stereo disparity on the left, and smaller disparity on the right. In inverted stereo, the left and right eye's views are switched, favoring perception of a nonrigid, truncated pyramid. (*f*) Left and right eye views showing the arrangement of the fixation cross and dots. A trial began only after the subject simultaneously perceived all four dots. (*g–j*) Illustration of an application of the *linear cue integration model* to the data of subject SW. (*g*) Weight of strength favoring the rigid depth isomer as a function of stereo disparity (in degrees of rotation between left- and right-eye images). (*h*) Weight of strength favoring the rigid depth isomer as a function of the log (base 10) of the ratio: (contrast of front)/(contrast of rear). (*i*) The model. Contrast strength W(k), disparity strength V(k), a (usually very small) rigidity bias B(k), and random noise ε are added; k indicates condition. A sum greater than 0 corresponds to the rigid Necker isomer being perceived; otherwise, the nonrigid isomer is perceived. (*j*) The result of the addition, illustrated for all 7 × 7 combinations of stereo (*g*) and front/rear contrast (*h*) cue values. (*k*) An energy map of perceived nonrigidity versus perceived structure illustrating two stable states of depth from motion (corresponding to the two perceptual depth isomers, m_1 and m_2). When a display is first turned on, a marble representing the current depth state is momentarily at the ridge between m_1, m_2, and moves according to a vertical force (gravity) and horizontal forces (generated by stereo and contrast cues) into one or the other energy well.

one of the energy wells, the lateral forces exerted by the steep walls (representing rapid loss of perceived rigidity away from the minimum) overwhelm the cue forces. The energy wells provide stable percepts even in the face of contradictory evidence.

This model offers a succinct representation of the fact that a quick acting force (such as line contrast) can exert a greater effect than stereopsis when displays move immediately on presentation, and why stereopsis is relatively much more effective when Necker cubes are first shown in a static (pre)view and only begin moving a second later. Stereopsis, which is perceived more slowly but which is a more powerful cue than the line contrast, wins when it is given ample time to exert itself in the static preview, but loses when line contrast tips the marble into a stable state before stereopsis can be effective.

Additive-Cue Models for the Perceived Depth of Surfaces

The additive framework has subsequently been applied by Johnston, Cumming, and Parker, (1993), Landy et al. (1991b), and Maloney and Landy (1989) to the problem of how the *shape* of recovered depth depends on various cues (stereo and texture, stereo and motion, texture and motion), which differ in depicted depths by small amounts. Although there are some anomalies in these data, it appears that the additive cue integration framework can account not just for the relative dominance of multistable percepts, but the depths recovered from non-ambiguous displays. In the context of an energy-surface conceptualization, these multiple cues combine with the motion cue to determine not just which minimum along an energy surface is selected, but some details of the shape of the energy surface as well.

Determining the Essential Stimulus Elements for Human Depth-from-Motion Processing

Experimental Methods in the Study of Depth-from-Motion

Introspection

Motion perception has played an important role in the history of psychology, especially among the Gestalt psychologists. They were much concerned with pure "phi motion"—a perception of motion between two briefly flashed bars that seemed to exist independently of any moving object. The introspective tradition continued with the discovery of the KDE by Wallach and O'Connell (1953), in the sense that emphasis was placed on introspective observations, such as judgments of amount of perceived depth, apparent coherence of moving points (resulting in one or more moving objects), apparent rigidity, and overall judgments of the quality of perceived depth. Two things were implicitly assumed in early KDE studies (i.e., Green, 1961): first, that kinetic depth is a unitary phenomenon so that any depth indicator would be sufficient as a yardstick and, second, that introspective judgments captured the essential features of the perception of depth from motion.

Objective Performance Tasks

With respect to the various measures of the kinetic depth effect, our own observations demonstrated that depthfulness, coherence, rigidity, and other properties were far from perfectly correlated, and no single such dependent variable could serve alone as an indicator of KDE (Dosher, Landy, and Sperling, 1989a). However, a more serious problem with introspective approaches is that they do not deal with the evolutionary purposes for which the ability to perceive depth from motion evolved. Evolution did not develop the systems subserving the recovery of 3D depth from 2D images in order to yield perceptions of rigidity or coherence or depthfulness, but in order to give organisms the ability to function in movement and action in the real world—to identify and discriminate shapes and surfaces in depth. For the study of KDE, it seemed to us essential to develop an objective task in which to study the capacity of depth-from-motion processes (Sperling et al., 1989, 1990). In our experiments, subjects are presented with kinetic depth displays, and asked to identify the shape from a lexicon of 55 similar shapes (figure 13.3).

What Can be Learned from Experiments with Feedback?

Given that objective measures of performance rather than introspection are used to study KDE, there remains the critical question of whether to provide to subjects feedback about the correctness or incorrectness of their responses. Feedback critically distinguishes what can be learned from experiments (Sperling et al., 1990): Experiments without feedback study the generalization of subjects' past experiences to the present experimental situation. In other words, experiments without feedback measure *achievement*; experiments with feedback can measure aptitude or ultimate *capacity* (Sperling et al., 1990). Aptitude and ultimate capacity are the words used to describe the asymptotic limit of a subject's performance (e.g., in a KDE shape identification task) after training with feedback.

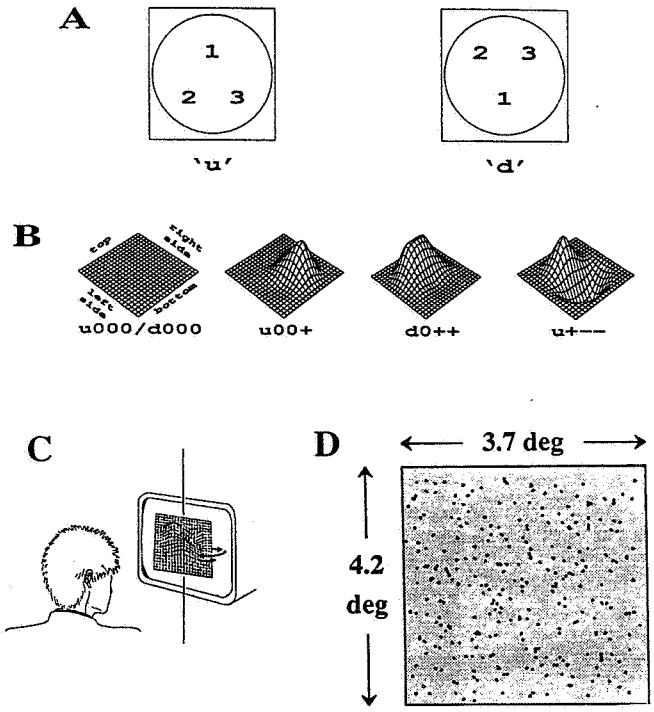

Figure 13.3
A 3D shape identification task. Subjects must indicate which shape they perceive from a 53-element lexicon of shapes built on splined bumps and depressions at three locations in one of the two configurations shown in (*A*). (*B*) Four sample shapes. (*C*) Random points on the surfaces of these shapes are projected and displayed undergoing sinusoidal rotation around a top-to-bottom axis (*B*). (*D*) A single freeze-frame of the projected points.

To prove that subjects require a particular cue (such as a motion flowfield) and cannot use other cues to solve a KDE shape-discrimination task, both an objective method and feedback are required. For example, when, in Sperling et al.'s (1989) procedure, subjects could not learn to use a texture density cue to solve the shape task, we know this failure represents an inherent biological limitation because subjects were given ample opportunity to learn.

Unfortunately, there is a hazard in experiments with feedback: subjects can use the feedback to learn to use incidental or artifactual cues (Braunstein and Todd, 1990). The solution to this potential problem is to refine the experimental procedure. Experiments with feedback intrinsically require more attention to detail and more experimenter work than experiments without feedback (Sperling et al., 1990).

Relative Depth Is Recovered from Motion Flow Fields

What is the evidence that relative depth information in kinetic depth displays is recovered from the motion flow field versus from a geometric calculation based on the 2D trajectories of identifiable features? To investigate the visual processes leading to recovery of 3D object depth, Dosher, Landy, and Sperling (1989b) began with kinetic depth displays and the objective shape identification task outlined in figure 13.3. 3D surfaces defined by dots were rocked back and forth 20° around a vertical axis. Dot density and contrast of the standard displays were set so as to yield shape identification performance in the high 90% range for most observers.

Changing 2D Dot Density—A Potential Confound

In a rigid object defined by dots painted on its surface, the 2D density of dots increases when the surface normal departs from the line of sight. In such surfaces, dot density is a cue to slant and thereby to depth. The density cue is eliminated by randomly adding or subtracting dots from any small area in which 2D dot density changes as a consequence of 3D motion so as to maintain locally uniform dot density. For motion confined to rocking of $\pm 20°$ around the line of sight, the elimination and addition of dots to maintain constant dot density involve about 5% of the dots per new frame. This percent of dots is small enough so the scintillation effect of adding and removing dots does not impair performance. On the other hand, when the density cue alone is presented in displays with motion cues removed, only one of three observers could use dot density to identify the shapes, and performance was 30%, compared to 90% performance with motion cues.

Dot Lifetimes Discriminate Flow Fields from Feature Tracking

To demonstrate that the necessary cue in their KDE displays was indeed motion and not the tracking of specific dots or groups of dots, Dosher et al. (1989b) used a dot lifetime manipulation in which each individual dot survived only for two frames (two frames is the minimum to define motion) and it was then replaced with another dot at a new location, which also survived for only two frames. On each frame (8 frames/sec), half the dots were replaced. The two-frame dot lifetime manipulation introduces an enormous amount of scintillation noise into the display plus much spurious motion "noise" produced by accidental apparent motion between unrelated dots. Nevertheless, KDE shape identification performance survives 2-frame lifetimes remarkably well. This result was confirmed by Todd and Bressan (1990) in subsequent, similar observations.

The 2-frame lifetime displays exemplify motion flow fields: they are devoid of larger features and even the microfeatures (hundreds of tiny dots) persist only for about 1/10 sec. The relatively good shape-identification performance in 2-frame lifetime displays demonstrates that the tracking of individual dots or groups of dots is not necessary for the perceptual recovery of depth from motion.

What Kind of Flowfield Is Necessary to Recover Depth-from-Motion?

First-Order Motion

A great deal is now known about visual motion analyzers as they apply to planar motion stimuli (van Santen and Sperling, 1984; Adelson and Bergen, 1985; Watson and Ahumada, 1985). It is quite well established experimentally that early motion analysis reflects a so-called first-order analysis of the stimulus—a computation based on the space–time Fourier motion components of the stimulus (van Santen and Sperling, 1984, 1985; Watson, Ahumada, and Farrell, 1986).

Second-Order Motion

In addition to the first-order analysis, the visual system is capable of detecting motion via a second-order analysis that requires initial space-time filtering followed by full-wave rectification prior to motion analysis (Chubb and Sperling, 1988, 1991). The evidence for second-order motion detection is the ability of subjects to detect motion in many kinds of displays that would be invisible to all the proposed first-order motion analyzers, the ability of subjects to perceive simultaneously first- and second-order motion embedded in the same display in opposite directions (Chubb and Sperling, 1989b; Solomon and Sperling, 1993), and the ability of subjects to discriminate the direction of first-order motion in the presence of strong second-order masking and vice versa (i.e., if there were only one system, the masking would destroy the ability of that system to perceive motion, Solomon and Sperling, 1994). Does the first- or second-order motion system subserve depth from motion?

Displays That Selectively Stimulate First- and Second-Order Motion-Analysis Systems

To study the dependence of depth from motion on first- and second-order motion detection mechanisms, Dosher, Landy, and Sperling (1989b) and Landy et al. (1991a) varied various aspects of KDE stimuli to make them relatively more or less useful to each of the motion detection systems. For example, normally all dots that define a surface were painted as white dots on a gray background. In alternating-polarity displays, the color of a moving dot alternated from white to black to white and so on in successive frames. Alternating polarity destroys the first-order motion signal but leaves the second-order signal completely intact. Alternating polarity was found to destroy subjects' ability to identify shape from motion in the displays, suggesting that first-order motion signals were necessary to solve the shape discrimination task.

In control experiments, subjects were required to judge the direction of motion of normal and alternating polarity displays. Even when normal and alternating-polarity stimuli were matched in this control task, the alternating-polarity stimuli failed to support depth from motion while the normal stimuli succeeded (Dosher et al., 1989b). Therefore, the failure of second-order stimuli to support depth from motion is not due to their inability to convey motion; it is a specific deficiency of second-order stimuli for a shape-identification task.

In another control experiment, subjects searched 3 × 3 arrays in which eight areas were defined by motion in one direction and one area was defined by motion in the opposite direction. Performance with second-order stimuli suffered as much in the search task as in the depth-from-motion task (Dosher et al., 1989b). However, in the search task, the origin of the problem was determined to be that subjects could search only one or two locations successfully for second-order motion. This is related to the more general observation that spatial resolution for second-order motion is much coarser in peripheral vision than in central vision (Solomon and Sperling, 1995); and since resolution is already poor in central vision, not enough remains to support performance in either depth from motion or in search tasks. While it is not obvious at how many locations and with what accuracy the motion flowfield needs to be sampled in order to solve the shape-from-motion task, it is obvious that the resolution of the second-order motion system was completely insufficient for Sperling et al.'s (1989) shape identification task.

Net Directional First-Order Motion Power (DP)

Alternating polarity is merely one of many stimulus transformations that selectively affect first-order versus second-order motion strength. Other transformations included alternating contrast strength in successive frames, interposing blank frames, replacing dots with other tokens (disks, lines, alternating-polarity point clusters), etc. The effects of all such stimulus transformations on

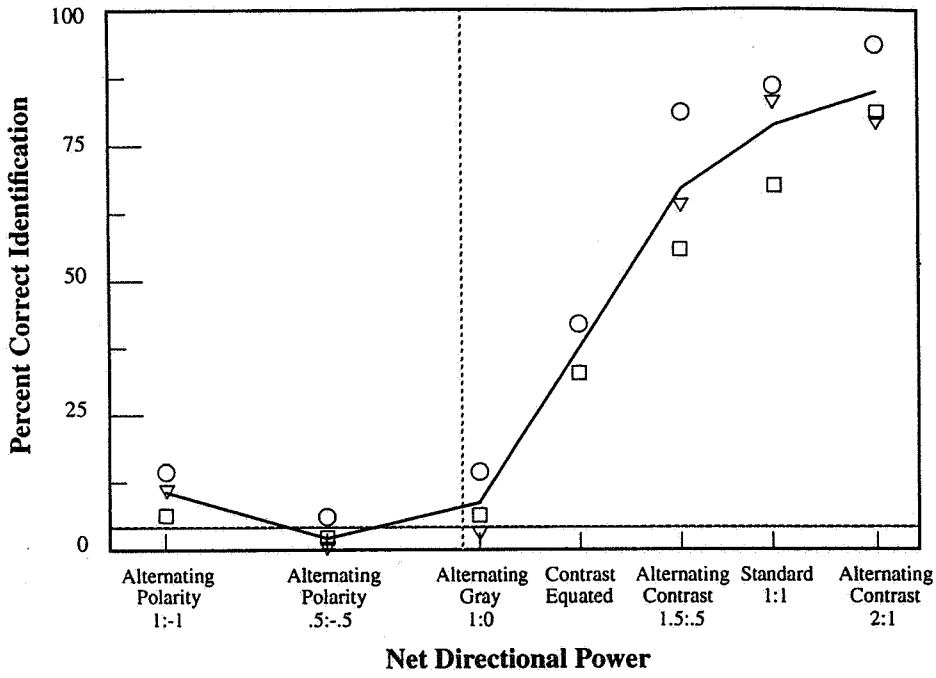

Figure 13.4
3D shape identification accuracy as a function of the *net directional power* (*DP*) of various stimulus types. Data are shown for three subjects. DP is computed from the Fourier amplitude transform of a stimulus trajectory (see text); it is the power of Fourier components above a threshold ε within a "window of visibility" (30 Hz × 30 c/deg).

subjects' accuracy in the shape-from-motion task are summarized in a single computation: net directional power. To derive directional power, Dosher et al. (1989b) compute the 2D Fourier power spectrum of the trajectory of a single dot. Following Watson, Ahumada and Farrell (1986), the analysis is confined to a window of visibility bounded by 30 cycles/degree and 30 Hz. Within this window, all Fourier components above a small threshold ε are given equal weight. Net directional power is simply the power of components in the intended direction minus the power in the opposite direction. Figure 13.4 shows that the proportion of correct shape-from-motion responses ranged from chance to very high levels in direct, monotonic relation to the directional power. In other words, the quality of the first-order motion signal directly predicts success in the shape-from-motion task.

Conclusions

1. The recovery of 3D depth from motion in monocular 2D displays reflects the output of a common depth channel that also records stereopsis.

2. The linear combination of cues to depth from motion is consistent with the computation of a bistable $(-1, +1)$ *motion–depth–sign*; the computation can be represented by an energy surface with two minima. The initial state is at the ridge separating the minima; conflicting cues represent forces directed towards opposite minima.

3. The sufficient cue to depth is the 2D motion flow field; subjects fail to derive depth from a changing texture-density cue nor do they require information derived from tracking specific dots or features.

4. Performance in 3D shape identification varies monotonically with the *net directional first-order motion power* contained in the stimulus.

Acknowledgments

This work was supported by ONR, Perceptual Sciences Program, grant N00014-88-K-0569 and by AFOSR, Visual Information Processing Program, grant 91-0178.

References

Adelson, E. H. (1985). Rigid objects that appear highly non-rigid. *Suppl. Invest. Ophthalmol. Visual Sci. 26, 56.*

Adelson, E. H., and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A 2, 284–299.*

Bennett, B. M., Hoffman, D. D., Nicola, J. E., and Prakash, C. (1989). Structure from two orthographic views of rigid motion. *J. Opt. Soc. Am A 6*, 1052–1069.

Biederman, I. (1987) Recognition-by-components: A theory of human image understanding. *Psychol. Rev 94*, 115–117.

Braunstein, M. L., and Todd, J. T. (1990). On the distinction between artifacts and information. *J. Exp. Psychol. Human Percept. Perform. 16*, 211–216.

Chubb, C., and Sperling, G. (1988). Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *J. Opt. Soc. Am. A Opt. Image Sci. 5*, 1986–2006.

Chubb, C., and Sperling, G. (1989a). Second-order motion perception: Space-time separable mechanisms. In *Proceedings: Workshop on Visual Motion* (March 20–22, 1989, Irvine, California) (pp. 126–138). Washington, D.C.: IEEE Computer Society Press.

Chubb, C., and Sperling, G. (1989b). Two motion perception mechanisms revealed by distance-driven reversal of apparent motion. *Proc. Natl. Acad. Sci. U.S.A. 86*, 2985–2989.

Chubb, C., and Sperling, G. (1991). Texture quilts: Basic tools for studying motion-from-texture. *J. Math. Psychol. 35*, 411–442.

Cornwell, H. G. (1976). Necker cube reversal: sensory or psychological satiation? *Percept. Motor Skills 43*, 3–10.

Dosher, B. A., Sperling, G., and Wurst, S. (1986). Tradeoffs between stereopsis and proximity luminance covariance. *Vision Res. 26*, 973–990.

Dosher, B. A., Landy, M. S., and Sperling, G. (1989a). Ratings of kinetic depth in multi-dot displays. *J. Exp. Psychol. Human Percept. Perform. 15*, 816–825.

Dosher, B. A., Landy, M. S., and Sperling, G. (1989b). Kinetic depth effect and optic flow: I. 3D shape from Fourier motion. *Vision Res. 29*, 1789–1813.

Green, B. F. (1961). Figure coherence in the kinetic depth effect. *J. Exp. Psychol. 62*, 272–282.

Grzywacz, N. M., and Hildreth, E. C. (1987). The incremental rigidity scheme for recovering structure from motion: Position vs. velocity based formulations. *J. Opt. Soc. Am. A 4*, 503–518.

Grzywacz, N. M., Hildreth, E., C., Inada, B. K., and Adelson, E. H. (1988). The temporal integration of 3-D structure from motion: A computational and psychopysical study. In W. von Seelen, G. Shaw, and U. M. Leinhos (Eds.), *Organization of Neural Networks: Structure and Models* (pp. 239–259). Weinhein, Germany: VCH.

Hildreth, E. C., Grzywacz, N. M., Adelson, E. H., and Inada, V. K. (1990). The perceptual buildup of three-dimensional structure from motion. *Percept. Psychophys. 48*, 19–36.

Johnston, E. B., Cumming, B. G., and Parker, A. J. (1993). Integration of depth modules: Stereopsis and texture. *Vision Res. 33*, 813–826.

Julesz, B. (1971). *Foundations of Cyclopian Perception*. Chicago: University of Chicago Press.

Koenderink, J. J., and van Doorn, A. J. (1986). Depth and shape from differential perspective in the presence of bending deformations. *J. Opt. Soc. Am. A 3*, 242–249.

Landy, M. S., Dosher, B. A., Sperling, G., and Perkins, M. E. (1991a). The kinetic depth effect and optic flow: II. First- and second-order motion. *Vision Res. 31*, 859–876.

Landy, M. S., Maloney, L. T., Johnston, E. B., and Young, M. (1991b). Measurement and modeling of depth cue combination: In defense of weak fusion. Mathematical studies in perception and cognition, 91–3. New York University Technical Report.

Longuet-Higgins, H. C., and Prazdny, K. (1980). The interpretation of a moving retinal image. *Proc. R. Soc. London B 208*, 385–397.

Loomis, J. M., and Eby, D. W. (1987). Perceiving 3-D structure from motion: Importance of axis of rotation. *Suppl. Invest. Opthalmol. Visual Sci. 28*, 234.

Maloney, L. T., and Landy, M. S. (1989). Psychophysical estimation of the human depth combination rule. In W. A. Perlman (Ed.), *Visual Communication and Image Processing IV, Proceedings of the SPIE, 1199*, 1154–1163.

Ono, H., and Steinbach, M. (1990). Monocular stereopsis with and without head movement. *Percept. Psychophys. 48*, 179–187.

Ono, M. E., Rivest, J., and Ono, H. (1986). Depth perception as a function of motion parallax and absolute-distance information. *J. Exp. Psychol. Human Percept. Perform. 12*, 331–337.

Orbach, J., Ehrlich, D., and Heath, H. (1963). A. Reversibility of the "Necker cube": I. An examination of the concept of "satiation of orientation." *Percept. Motor Skills 17*, 439–458.

Pentland, A. (1989). Part segmentation for object recognition. *Neural Comp. 1*, 82–91.

Pentland, A., and Sclaroff, S. (1991). Close form solutions for physically based shape modelling and recognition. *IEEE Transact. Pattern Anal. Machine Intelligence 13*, 715–729.

Schwartz, B. J., and Sperling, G. (1983). Luminance controls the perceived 3D structure of dynamic 2D displays. *Bull. Psychon. Soc. 21*, 456–458.

Solomon, J. A., and Sperling, G. (1994). Fullwave and halfwave rectification in motion perception. *Vision Res. 34*, 2239–2257.

Solomon, J. A., and Sperling, G. (1995). 1st- and 2nd-order motion and texture resolution in central and peripheral vision. *Vision Research, 35* (1). (In press.)

Sperling, G. (1970). Binocular vision: a physical and a neural theory. *Am. J. Psychol. 83*, 461–534.

Sperling, G., and Dosher, B. (1987). Predicting rigid and nonrigid perceptions. *Suppl. Invest. Ophthalmol. Visual Sci. 28*(3), 362.

Sperling, G., Dosher, B., and Landy, M. S. (1990). How to study the kinetic depth effect experimentally. *J. Exp. Psychol. Human Percept. Perform. 16*, 445–450.

Sperling, G., Landy, M. S., Dosher, B., and Perkins, M. (1989). The kinetic depth effect and identification of shape. *J. Exp. Psychol. Human Percept. Perform. 15*, 826–840.

Spitz, H. H., and Lipman, R. S. (1962). Some factors affecting necker cube reversal rate. *Percept. Motor Skills 15*, 611–625.

Thurstone, L. L. (1947). *Multiple-Factor Analysis: A Development and Expansion of the Vectors of Mind*. Chicago: University of Chicago Press.

Todd, J. T., and Bressan, P. (1990). The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Percept. Psychophys. 48*, 419–430.

Ullman, S. (1979). *The Interpretation of Visual Motion*. Cambridge, MA: MIT Press.

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion. *Perception 13*, 225–274.

van Santen, J., and Sperling, G. (1984). A temporal covariance model of motion perception. *J. Opt. Soc. Am. A 1*, 451–473.

van Santen, J., and Sperling, G. (1985). Elaborated Reichardt detectors. *J. Opt. Soc. Am. A 2*, 300–321.

Wallach, H., and O'Connell, D. N. (1953). The kinetic depth effect. *J. Exp. Psychol. 45*, 205–217.

Watson, A. B., and Ahumada, A. J. (1985). Model of human visual-motion sensing. *J. Opt. Soc. Am. A 1*, 322–342.

Watson, A. B., Ahumada, A. J., and Farrell, J. E. (1986). Window of visibility: A psychophysical theory of fidelity in time-sampled visual motion displays. *J. Opt. Soc. Am. A 3*, 300–307.