Journal of Experimental Psychology:
Human Perception and Performance
1985, Vol. 11, No. 6, 689-710

0096-1523/85/$00.75

Perception of Structure From Motion: Is Projective Correspondence of Moving Elements a Necessary Condition?

James T. Todd
Brandeis University

A fundamental assumption of almost all existing computational analyses of the perception of structure from motion is that moving elements on the retina projectively correspond to identifiable moving points in three-dimensional space. The present investigation was designed to determine the psychological validity of this assumption in several different contexts. The results demonstrate that the ability of human observers to perceive structure from motion is much more general than would be reasonable to expect on the basis of existing theory. Observers can experience a compelling kinetic depth effect even when the pattern of optical motion is contaminated by large amounts of visual noise (e.g., where the signal to noise ratio is less than 0.15). Moreover, the optical deformations of shading, texture, or self-occluding contours, which would be treated as noise by existing computational models, are analyzed by human observers as perceptually salient sources of information about an object's three-dimensional form. These results suggest that the modular analyses of visual information that currently dominate the literature will have to be modified if they are to account for the high level of generality exhibited by human observers.

Many different sources of visual information are potentially available to human observers about the three-dimensional structure of the environment. One of the most perceptually salient, however, is the pattern of optical change produced by a moving object in space. The importance of motion for the perception of three-dimensional form was first demonstrated over 30 years ago in a classic series of experiments by Wallach and O'Connell (1953). In what is now known as the *kinetic depth effect*, these authors showed that a moving visual image can provide perceptually salient information about the three-dimensional form of an object, even though a static image of the same object is perceived as two-dimensional. Much has been learned about the kinetic depth effect in the three decades since its initial discovery (see Braunstein, 1976, for an excellent review). Some researchers in recent years have even proposed specific computational mechanisms that could potentially provide a formal theoretical explanation of how the three-dimensional form of an object could be reliably determined from its pattern of projected motion (e.g., Koenderink & van Doorn, 1975, 1977; Lee, 1974; Longuet-Higgens & Prazdny, 1981; Todd, 1981, 1982; Ullman, 1979).

A fundamental problem for any formal analysis of the perception of structure from motion is that a moving visual image is mathematically ambiguous—that is, there are an infinite number of object transformations in three-space that are projectively equivalent to any given moving image on a two-dimensional display surface. Most theorists have addressed this problem by postulating constraints on the structure of the environment that limit the number of possible three-dimensional interpretations to be considered. The constraints imposed by existing analyses include restrictions on the type of object that can be analyzed (e.g., smoothly curved surfaces), the nature of its motion (e.g., rotation about a fixed axis), and the viewing distance from which it is observed (i.e., parallel or polar projection). Different analyses tend to have different limitations because of variations in their underlying assumptions. Indeed, a considerable amount of psychophysical research has been directed

I am most grateful to Ennio Mingolla, John Pittenger, and William Warren for their collaborative efforts during various phases of this research, and to Joseph Lappin for his helpful comments on an earlier draft of the manuscript.

Requests for reprints should be sent to James T. Todd, Department of Psychology, Brandeis University, Waltham, Massachusetts 02254.

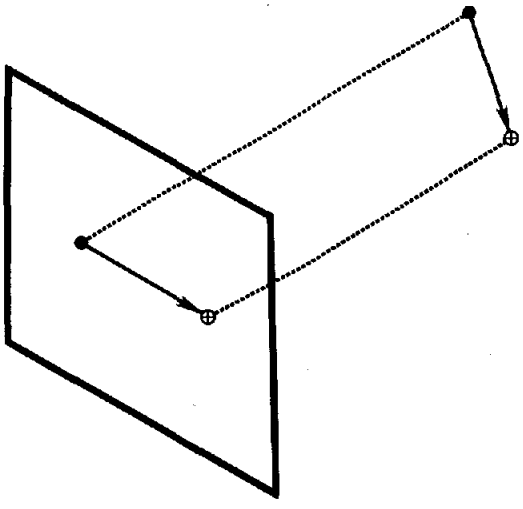

Figure 1. The optical projection of a single point moving in three-dimensional space. (The projective relation between the position of the point in three-space and its corresponding position in the picture plane is represented with a dotted line. The open and filled circles are used to designate the positions of the point at different moments in time, and the resulting displacements in both the picture plane and in three-space are represented by solid arrows.)

at determining which limitations correspond most closely to the perceptual capabilities of actual human observers (e.g., Braunstein & Andersen, 1985; Doner, Lappin, & Perfetto, 1984; Lappin, Doner, & Kottas, 1980; Lappin & Fuqua, 1984; Todd, 1981, 1982, 1984).

Despite these differences among existing analyses, there is at least one important characteristic that they all share in common—namely, the assumption that two-dimensional movements of elements on the retina (or on a visual display surface) are the optical projections of identifiable moving elements in three-dimensional space (see Figure 1). Although this assumption of projective correspondence is seldom stated explicitly in most computational analyses, it is of vital importance to their successful application. If, for example, the correspondence assumption were violated as shown in Figure 2, then the resulting sequence of images could not be given a correct three-dimensional interpretation.

The issue of projective correspondence is easily overlooked in most psychophysical investigations because the usual methods of stimulus generation ensure that the correspondence assumption can always be satisfied. That is not the case, however, when dealing with natural images. Consider, for example, the surface of a lake on a windy day. The choppiness of the water gives the surface a textured appearance, but the individual texture elements seem to appear and disappear at random without producing a well-defined optical motion. A similar problem arises when dealing with moving images of smoothly curved surfaces that contain shadows, specular highlights, or self-occluding boundaries. The image contours produced by these phenomena will be deformed over time, but the resulting deformations will not correspond to the movements of an identifiable locus of points in three-dimensional space.

These examples demonstrate that there are visual events encountered in nature for which the assumption of projective correspondence is invalid. Existing computational analyses for determining structure from motion are of little use in these situations, but is human perception constrained in the same way? The research described in the present article was specifically designed to address this question. The perceptual significance of violations of the correspondence assumption were examined in three

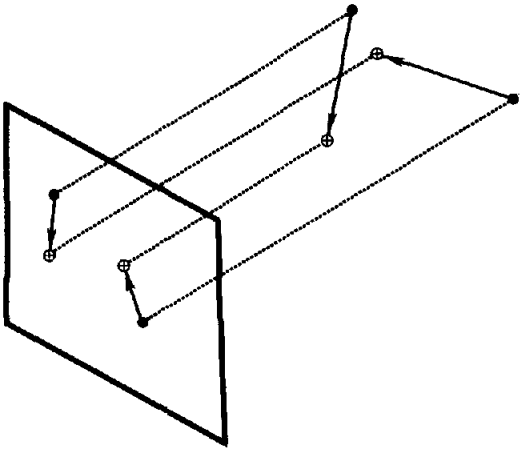

Figure 2. The optical projection of two moving points that are inappropriately matched. (That is to say, the projection of one point at a particular moment in time is matched with the projection of some other point at a subsequent moment, and vice versa. On the basis of current theory, it should not be possible to analyze structure from motion if the optical elements at different moments in time are inappropriately matched.)

different contexts: (a) the movements of configurations of points, (b) the deformations of continuous contours, and (c) the deformations of smoothly varying patterns of shading or texture. The results of these experiments suggest that the perceptual capabilities of human observers are much more general than one would expect on the basis of current models.

Movements of a Configuration of Points

Consider a point-light display in which a sequence of images is presented in rapid succession. A fundamental problem for the computational analysis of this type of display is to match up the points in subsequent images in order to define a two-dimensional displacement for each individual element.[1] The primary goal of this matching process is to satisfy the correspondence assumption. That is to say, two points should be matched only if they are projectively related to a single identifiable point in three-dimensional space (e.g., see Figures 1 and 2). It is important to keep in mind that the usual procedures for generating a point-light display through computer simulation guarantee that a correct matching configuration will always exist for any pair of images in a given sequence. A serious problem arises, however, if a display is contaminated by visual noise. Any displacement involving a noise element that is detected during the matching process will fail to satisfy the correspondence assumption and could therefore impair any subsequent computations of a moving object's three-dimensional structure.

My interest in the problem of how visual noise affects the perception of structure from motion began several years ago in a series of conversations with my colleague John Pittenger (personal communication, 1979). In an unpublished experiment, Pittenger attempted to show that a change in a higher order variable, such as a texture gradient, could provide information about an object's motion even when there is no motion of the individual texture elements. To test this hypothesis he created a series of motion picture sequences of a moving tray of rice. Between each frame in a sequence he shook up the rice to obtain a new random pattern. Thus, there was no correspondence between individual elements, although the density of texture varied smoothly with changes in the depth or orientation of the surface. The outcome of this clever and arduous procedure turned out to be disappointing. The resulting stimulus displays bore little resemblance to a moving surface, suggesting that correspondence of elements over time may indeed be a necessary condition for the perception of motion.

This conclusion was reinforced in a similar experiment by Lappin, Doner, and Kottas (1980). They showed observers two-frame apparent motion sequences depicting 512 randomly positioned points on a rotating sphere. The three-dimensional structure of these displays was easily detected if there was perfect correspondence between the two separate images, but performance deteriorated dramatically with the introduction of even a small amount of visual noise. However, other research by Petersik (1979) and Doner, Lappin, and Perfetto (1984) has demonstrated that the perception of structure from motion can tolerate very large amounts of noise (e.g., with signal to noise ratios less than one) if the stimulus displays include longer sequences with more than two images.[2]

The present series of experiments began with a variation of these earlier investigations. Observers were required to estimate the rotation in depth of a planar surface for visual displays in which the level of correspondence was systematically manipulated.

Experiment 1

Method

Subjects. Three observers, including the author and 2 Brandeis graduate students, participated in the experiment.

---

[1] Adelson and Bergen (1985) have recently described a method of detecting motion from sequences of static images that does not require the matching of individual elements. In their approach the velocity of motion in local regions of the retina is determined directly from the sampled pattern of spatiotemporal energy. These local velocities would still have to satisfy the assumption of projective correspondence, however, in order to be compatible with existing methods for computing structure from motion.

[2] The signal-to-noise threshold for detecting coherent motion in the image plane is apparently much lower than the threshold for perceiving structure from motion. In a remarkable series of experiments by van Doorn and Koenderink (1982a, 1982b, 1982c, 1983) it has been demonstrated that human observers can detect patterns of dots translating in the image plane with signal-to-noise ratios as low as 0.01.

Neither of the graduate student volunteers was familiar with the theoretical issues being investigated or the specific details of how the displays were generated.

Apparatus. The stimuli were produced using an LSI-11/23 microprocessor and displayed on a Terak 8600 color graphics system at a viewing distance of approximately 50 cm. Head and body movements were not restricted. The stimuli were presented within a rectangular window of the display screen that was 18 cm along the vertical axis and 25 cm along the horizontal axis. The spatial resolution within this viewing window was 320 × 240 pixels. Thus, each pixel had a horizontal and vertical extent of approximately 0.075 cm, producing a visual angle of approximately 5 min.

Stimuli. Observers were presented with computer simulations of planar surfaces rotating in depth at a simulated viewing distance of 50 cm. At the beginning of each display, the surface would be oriented at some angle relative to the display screen; it would rotate to a vertical orientation and then return to its initial position. The changes in slant angle were varied sinusoidally by stepping through a sequence of 24 frames at a rate of ½ Hz for seven complete oscillation cycles. Each of the simulated surfaces contained 100 luminous blue dots that were distributed at random with uniform probability density. All of these dots were visible at the most extreme slant angle, but because they could be occluded by the edges of the display screen, some of the dots at the top and bottom would disappear during other portions of the oscillation cycle.

At each frame transition in a display, some of the dots (called *noise elements*) were randomly repositioned on the simulated surface. The remaining dots (called *signal elements*) maintained their original positions. The percentage of signal elements determining the level of correspondence could vary from 0% to 100%. The specific designation of whether an element was signal or noise could change at random from one transition to the next, but the overall level of correspondence always remained constant within a given display. Thus, at a 50% level of correspondence each element had a probability of 1/2 of maintaining its position in two successive frames and a probability of 1/8 of maintaining its position in four successive frames. At a 10% level of correspondence these probabilities would be reduced to 1/10 and 1/1,000, respectively.

The observers' task in the experiment was to estimate the amplitude of oscillation (i.e., how far the surface rotated in depth). The possible responses included 0°, 10°, 20°, 30°, 40°, and 50°. (Zero-degree oscillation angles were not used in any of the simulations, but were included as a response category because some of the displays were not perceived as rotating.) There were five possible angles of rotation used in the simulations (10°, 20°, 30°, 40°, and 50°) and five possible levels of correspondence (0%, 12%, 25%, 50%, and 100%). These were presented in all possible combinations.

Procedure. Before an observer saw any of the computer-generated displays, the experimenter described verbally what they would depict. The observers were told that they were to judge the perceived rotation in depth of a planar surface. Next they were shown a diagram depicting the six possible rotation angles included in the response categories. This diagram remained in view throughout the experiment. The observers were instructed to indicate the perceived rotation angle of each display by pressing the appropriate key (0–5) on the computer keyboard.

All of the observers participated in two experimental sessions. During each session a randomized sequence of the 25 displays (5 rotation angles × 5 levels of correspondence) was presented five times in succession. The experimenter stayed in the room during the entire first pass to answer any questions. All of the data from this pass were treated as practice and excluded from subsequent analyses. An experimental session took approximately 40 min. No feedback of an observer's performance was given until after both sessions were completed.

Results and Discussion

Figure 3 shows the mean rotation estimates plotted against the simulated rotation for all five levels of correspondence. As is evident in the figure, the observers' judgments were almost perfectly accurate when the level of correspondence was 100% (see also Flock, 1964; Gibson & Gibson, 1957). When the level of correspondence was less than 100%, however, the observers tended to underestimate the rotation angles. This underestimation was particularly severe in the 0% correspondence condition. In that case, the observers perceived virtually no rotation at all even with the highest rotation angles. Figure 4 shows the average error (collapsed over rotation angles) as a function of the level of correspondence. It is clear from this figure that some amount of correspondence is required to perceive rotation in

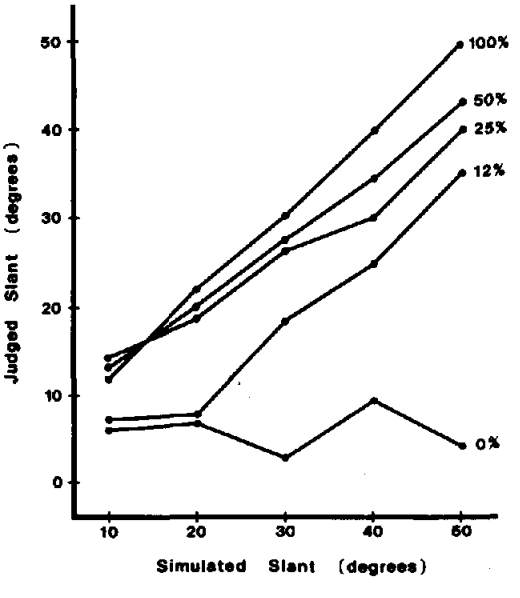

Figure 3. The mean rotation estimates of 3 observers as a function of simulated rotation for the five levels of correspondence in Experiment 1.

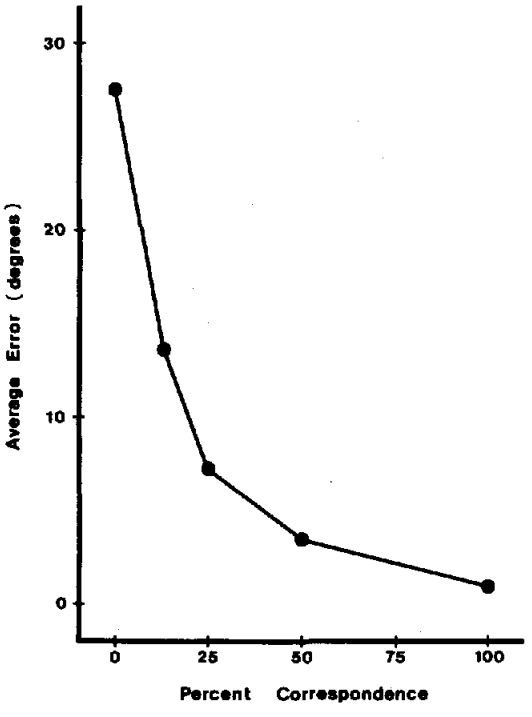

Figure 4. The average error in Experiment 1 as a function of the level of correspondence.

depth but that the level of correspondence need not be large (see also Petersik, 1979). A level of only 12% is sufficient for reasonably accurate judgments.

These findings are all consistent with the observers' phenomenal impressions of the displays. With a perfect 100% level of correspondence, a display appears as a rigid surface rotating in depth; at an intermediate level of correspondence the rotating surface appears to be scintillating; but with no correspondence at all a display looks more like a swarm of flies than like any type of coherent surface.

It is important to keep in mind that there was potential information available about the rotating surfaces in the 0% correspondence condition. The gradient of texture density varied systematically throughout the oscillation cycle, but as Pittenger (personal communication, 1979) discovered in his earlier experiment, observers are apparently unable to make use of that information for perceiving rotation in depth. It should also be noted, however, that there are forms of texture information besides density (e.g., size or shape gradients) that are more perceptually salient (e.g., see Cutting & Millard, 1984). It remains to be demonstrated whether changes in these other gradients might be an adequate stimulus for the perception a rotating surface.

One possible explanation of the observers' performance in this experiment is based on the distinction between signal and noise elements. Because the rotation angle between successive frames in a display was never more than a few degrees, the projected displacement of any given signal element would be relatively small (i.e., most were in the range of 5 to 15 min of visual angle). The projected displacement of a noise element, in contrast, would be many times larger because of its repositioning. Moreover, because the surfaces were sparsely populated with dots, it was highly improbable that a noise element would be repositioned near where a dot had been located in the preceding frame (i.e., an average noise element would be separated by more than 1° from its nearest neighbor—see Figure 5). In other

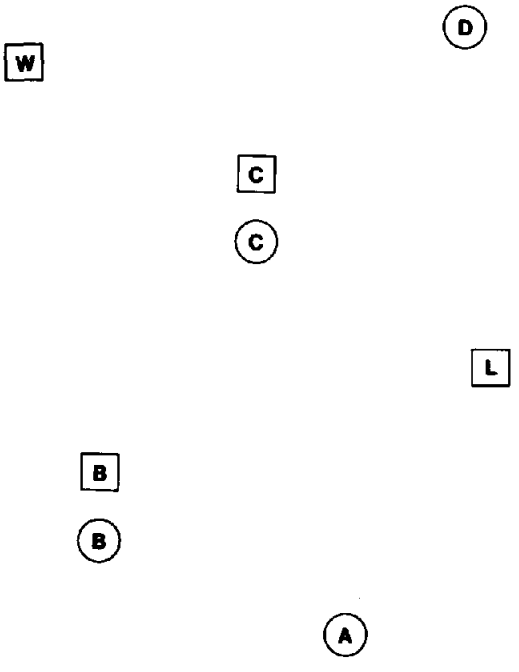

Figure 5. A possible configuration of elements over two consecutive images in a local region of a display with 50% correspondence. (The letters in the figure are used to identify individual elements, and the circles and squares are used to distinguish the two separate images. Note that the corresponding [signal] elements are relatively close together, whereas the noncorresponding [noise] elements are relatively far apart.)

words, if the projection of an element in one frame was close to where an element had been projected in the preceding frame, there was a high probability it was a signal element. All of this suggests that the observers may have been able to distinguish the signal elements from the noise elements based on their projected displacements and that only the signal elements were used in estimating the rotation angles. Such a strategy would be expected to produce a disproportionate number of errors when the level of correspondence is close to zero, which is exactly what occurred in the actual experiment.

This type of strategy for segregating signal from noise could be implemented physiologically by having a population of motion detectors that are able to function as an autonomous unit and that are all tuned within a limited range of velocities. In a recent series of experiments, van Doorn and Koenderink (1982a, 1982b, 1982c, 1983) have provided strong psychophysical evidence that such populations do indeed exist within the human visual system. They have also pointed out, moreover, that functionally autonomous populations of motion detectors are needed to account for transparency effects where two or more moving surfaces are perceived simultaneously at the same place in the visual field (e.g., see Gibson, Gibson, Smith, & Flock, 1959).[3]

If the signal and noise elements in the present displays are distinguished by the magnitudes of their projected motion, then there are two types of manipulations that ought to have a severe impact on observers' perceptions. If, for example, the angular displacements between successive frames were increased, then the projected displacements of the signal elements could be increased to the point where they would no longer be distinguishable from the noise elements. This manipulation was performed by Doner et al. (1984), and, as expected, it produced a dramatic drop in the perceived coherency of their displays. A similar effect should also be possible by increasing the density of elements on a surface. In that case the average distance between noise elements in successive frames could be reduced to the point where they would again be indistinguishable from the signal elements. Demonstration 1 was designed to test this prediction.

Demonstration 1

Method

The apparatus and general procedure were roughly equivalent to those used in Experiment 1. Two displays were created of a planar surface rotating back and forth in depth through a 50° angle. Each display contained 10,000 points, which were distributed at random on the simulated surface with uniform probability density. In one of the displays there was a perfect 100% correspondence between successive frames in the sequence, whereas in the other, the level of correspondence was reduced to 50%. The density of texture in this display was sufficiently large so that the majority of noise elements would be repositioned to a new location that was no more than one or two pixels from the position of an element in the preceding frame. (The width of a pixel was approximately 5 min of arc.)

A wide variety of observers have viewed these displays in both a laboratory setting and in public presentations. In each case the observers have been asked to comment on the perceived rotation in depth of the simulated surfaces.

Results and Discussion

The increase in texture density of these displays relative to those used in Experiment 1 has a clear-cut effect on perceptual experience. For the display with 100% correspondence, the effect of this increased density is positive. That is to say, the perception of rotation in depth for a densely textured surface with 10,000 points is considerably more compelling than that of a similar surface with only 100 points. For the display with 50% correspondence, however, the effect of density is reversed. In that case, a one hundredfold increase in the number of points reduces the impression of rotation in depth. Nevertheless, although the perception of structure from motion in the presence of noise is attenuated by large increases in texture density, it is by no means eliminated. Most observers insist that they are able to detect a rotating surface even though it appears to be camouflaged by the pattern of scintillation.

---

[3] An interesting phenomenological aspect of the present displays that is not readily explained by van Doorn and Koenderink's hypothesis is that the pattern of scintillation appears to be attached to the rotating surface. This occurs, however, only when the signal elements are selected randomly at each frame transition. For other displays in which the set of signal elements remains constant throughout the entire sequence, the perceived surface appears to rotate through a transparent vertical plane of scintillating noise.

In the discussion of Experiment 1 it was suggested that if signal and noise elements are distinguished by the magnitudes of their projected displacements, then the perception of structure from motion should be severely impaired by an appropriate increase in texture density. Although the present demonstration confirms this prediction, it also poses a new problem: How could an observer perceive any coherent motion at all in such a densely textured display with only 50% correspondence? Because any of the known methods of computing structure from motion would be completely overwhelmed by such a high level of noise, it seems reasonable to speculate that some other process of noise reduction might also be at work in this situation. Experiment 2 was designed, therefore, in an effort to explore this process in greater detail.

Experiment 2

Method

The procedure was basically the same as in Experiment 1. The primary difference was that the stimuli were generated with two types of noise. For some of the displays, the noise elements were unconstrained as in the previous experiment—that is to say, they could be randomly repositioned anywhere on the surface. For other displays, however, a different type of constrained noise was employed. A noise element in that case would be repositioned a short distance from its previous location in a randomly selected orientation. (The length of this displacement was 0.2 cm, which was approximately 15 min of arc.) Another important difference from the previous experiment was that no element was allowed to maintain its position for more than two successive frames in either noise condition. The displays could be generated with two possible levels of correspondence (0% and 50%) for each type of noise, or with no noise at all at a 100% level of correspondence. These could occur with five different rotation angles (10°, 20°, 30°, 40°, and 50°) in all possible combinations. The same 3 observers who participated in Experiment 1 judged all 25 displays eight times over a period of two sessions. No feedback was given until after the experiment was completed.

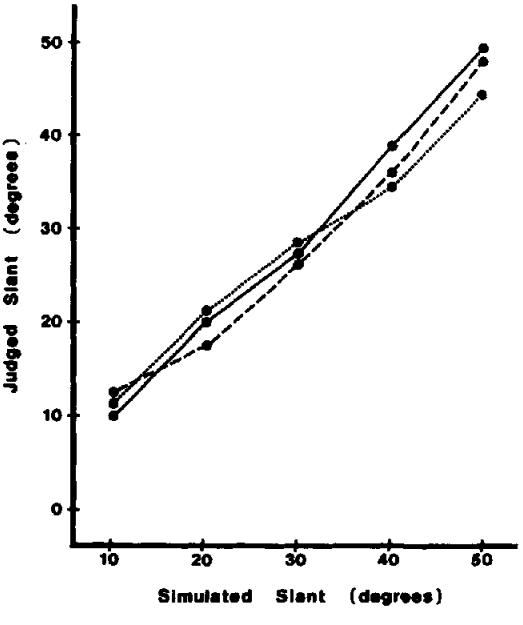

Figure 6. The mean rotation estimates of 3 observers as a function of simulated rotation for the constrained noise conditions of Experiment 2. (The 0%, 50%, and 100% correspondence conditions are represented by dotted, dashed, and solid lines, respectively.)

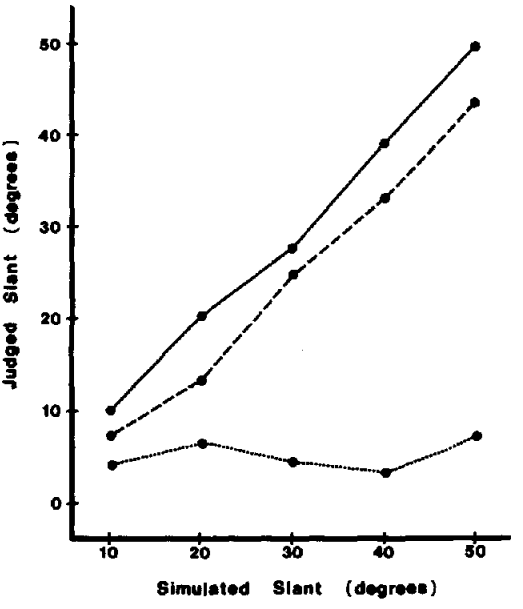

Figure 7. The mean rotation estimates of 3 observers as a function of simulated rotation for the unconstrained noise conditions of Experiment 2. (The 0%, 50%, and 100% correspondence condition are represented by dotted, dashed, and solid lines, respectively.)

Results and Discussion

The data for the constrained and unconstrained noise are given in Figures 6 and 7, respectively, where the observers' rotation estimates are plotted against the simulated rotation for each level of correspondence. (The data for the 100% correspondence condition

is repeated in both graphs.) It is important to note in these figures that the observers could perceive rotation in depth despite the fact that no element maintained its position for more than two consecutive frames (cf. Ullman, 1979). Indeed, the results for the unconstrained noise condition were virtually identical to those obtained in Experiment 1. With 0% correspondence no rotation was perceived, but with 50% correspondence there was a high level of performance. The results obtained with the constrained noise were quite different, however. The level of performance in that case remained high even in the 0% correspondence condition. This is revealed most clearly in Figure 8, which shows the average error (collapsed over rotation angles) as a function of correspondence for each type of noise.

The constrained noise condition was specifically designed so that the signal and noise elements could not be distinguished on the basis of their projected displacements, yet the observers' judgments in this condition were

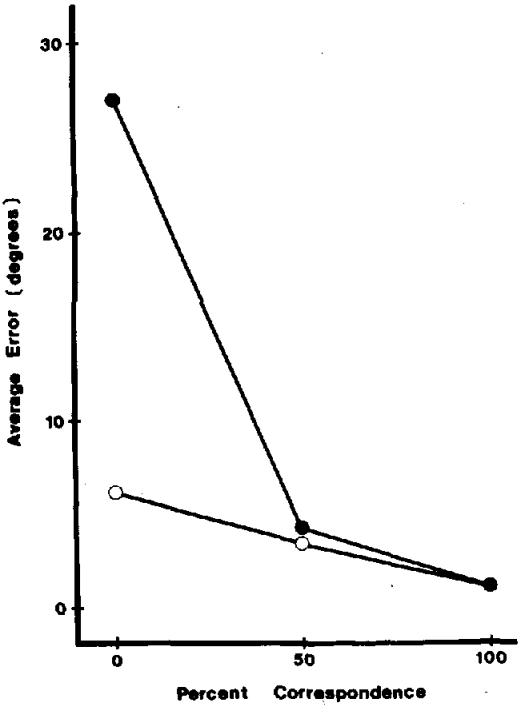

Figure 8. The average error in Experiment 2 as a function of the level of correspondence. (The constrained and unconstrained noise conditions are represented by open and closed circles, respectively.)

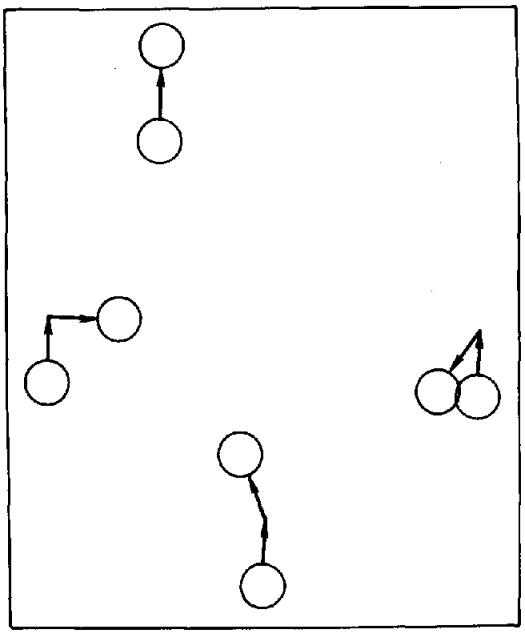

Figure 9. A possible configuration of elements over two consecutive images in a local region of a display with constrained noise. (As shown in the figure, the projected displacement of each noise element is a combination of two vectors: one due to the rotation of the surface and the other due to the repositioning of the element on the surface. The rotation component is constant within a given local region, whereas the repositioning component varies at random from one element to the next.)

significantly more accurate than with the unconstrained noise condition at low levels of correspondence. What type of process could account for this high level of performance? A key difference between the constrained and unconstrained conditions is that the constrained noise elements in one frame could be matched with their corresponding elements in the next. The constrained noise elements in successive frames could be matched because of their small projective displacements—the same reason they could not be distinguished from the signal elements. It is important to keep in mind, moreover, that the projected displacement of each noise element was a combination of two vectors: one due to the rotation of the surface, the other due to the repositioning of the element on the surface (see Figure 9). If several displacement vectors within a local neighborhood were averaged together, the repositioning components would

tend to cancel out each other, and the resulting average would closely approximate the shared rotational component.

An optimally efficient averaging process would be performed over several spatial scales simultaneously. This would allow an observer to perceive the global motion of a surface as well as the relative local movements of individual elements across the surface. This is consistent with the observers' subjective reports in the present experiment. All of them agreed that the displays with constrained noise appeared as a rotating surface on which a large number of luminous ants were crawling around at random. It should also be noted that signal averaging can be performed over regions of time as well as space. Indeed, the results of Doner et al. (1984) provide strong psychophysical evidence that some sort of temporal averaging may be involved in the perception of structure from motion.

All of this suggests that the human visual system has at least two different strategies for dealing with noisy inputs: In some situations it may be possible to separate signal from noise by isolating populations of moving elements based on the magnitude of their optical motions. In others, it may be possible to overcome the effects of noise by averaging optical motions within local regions of space and/or time. It is important to point out that neither of these strategies for coping with noise is necessarily incompatible with a computational analysis based on an assumption of projective correspondence. Although the immediate inputs of such an analysis must still satisfy the correspondence assumption, it is possible that moving elements on the retina may violate this assumption if an intermediate process such as signal averaging eliminates those violations prior to the analysis of structure from motion.

Deformations of Continuous Contours

A second common situation where the assumption of projective correspondence can run into difficulty occurs when a visual image contains continuous contours that deform over time. Consider, for example, a flat circular disk that is rotating in depth. The optical projection of this disk in the picture plane will be gradually transformed from a circle, through a series of ellipses of varying eccentricity to a straight line. This sequence will repeat itself, first in one direction, then in reverse for as long as the disk continues to rotate. How would one compute structure from motion in this context? It is important to keep in mind that most existing analyses are specifically designed to compute the three-dimensional structure of an array of points from the projected displacements (or velocities) of those points on a visual projection surface. In order to satisfy the assumption of projective correspondence, it is necessary to identify the optical projections of a given point at different moments in time, or, equivalently, to determine the projected velocity of a point at an instantaneous moment in time. The solution to this problem is especially difficult when dealing with continuous contours. In the case of apparent motion, one cannot track the displacement of a given point because all points along the contour are indistinguishable. Similarly, in the case of continuous motion, one cannot determine the component of velocity parallel to the contour at any given location (see Hildreth, 1983).

A number of different techniques have been proposed in the literature that could conceivably be used for overcoming these difficulties, either by imposing additional constraints so that a single velocity can be assigned at each point along a contour (e.g., Hildreth, 1983) or by analyzing an object's structure directly from the pattern of deformation (e.g., Koenderink & van Doorn, 1977). For these analyses a higher order variation of the correspondence assumption is applicable. That is to say, it is assumed that moving contours on the retina (or in a visual display) are the optical projections of an identifiable *locus* of points moving in three-dimensional space.

Unfortunately, there is a wide variety of image contours produced by shadows, specular highlights, or the self-occluding boundaries of smooth surfaces that do not satisfy this assumption. Consider the case of a solid object that is bounded by a smooth homogeneous surface. At each point on the object we can define two vectors: one that is normal to the surface, and another that is oriented toward the point of observation. The optical contour that bounds the object's projection will correspond to a locus of points for which these vectors are perpendicular (i.e., the contour is formed by the visual rays that just graze the

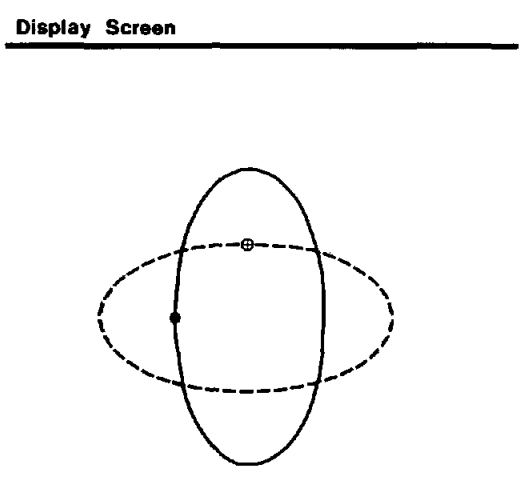

Figure 10. Two images of a rotating ellipsoid as viewed parallel to the display screen from below. (The solid line depicts the outline of the ellipsoid at one moment in time when its major axis is perpendicular to the display screen, and a single identifiable point on the surface is designated by a filled circle. The dashed line represents the same ellipsoid after a rotation of 90°, and the identifiable point in that case is represented by an open circle.)

object's surface). If the object moves, this optical contour will be deformed, but the locus of surface points to which it corresponds will also be continuously changing (see Figures 10 & 11).

The inability of existing computational models to deal with such severe violations of the correspondence assumption was clearly recognized by Marr (1982):

> This point is important. For example, failure to recognize it held Wallach and O'Connell (1953) up for years by their own admission. They could not understand why the shadow of a bent wire should be different from the shadow of a smooth solid object. If a wire is rotated, its shadow moves, and one instantly perceives the wire's three-dimensional shape; if a solid object is rotated, its shadow moves but one cannot perceive its shape. The reason is that the shadow of the wire produces an outline that is effectively in one-to-one correspondence with fixed points on the wire, each having a definite physical location that changes from frame to frame, admittedly, but that always corresponds to the same piece of wire. For the rotating object this is just not true. From moment to moment, the points on the silhouette correspond to quite different points on the object's surface. The image primitives are no longer effectively tied to a constant physical entity. Hence the shape recovery process fails. (p. 105)

An obvious implication of Marr's argument is that a psychologically valid model of the analysis of structure from motion need not be concerned with the deformations of self-occluding contours because those deformations are an inadequate stimulus for human perception. The evidence to support this claim is rather weak, however. In an aside comment of their original article, Wallach and O'Connell (1953) did indeed state that smoothly curved objects do not produce a compelling kinetic depth effect, but they provided no details whatsoever of the specific manipulations by which they reached this conclusion. As it turns out, the conclusion is incorrect, as will be shown in the following demonstration.

Demonstration 2

Method

Four different computer simulations were generated of solid objects rotating in three-dimensional space. In Display 1, the simulated object was a horizontal ellipsoid with x-y-z semiaxes of 7.8, 2.9, and 2.9 cm, respectively, in its initial orientation. The center of this object was located in the center of the display screen. In Display 2, the simulated object was a vertical ellipsoid with x-y-z semiaxes of 1.0, 4.9, and 1.0 cm, respectively. The center of this object in its initial position was located 7.8 cm to the right of the center of the display screen. In Display 3, these two objects were presented together as a surface of intersection. Finally, in Display 4, the two objects were again presented together but were displaced vertically so that their optical projections did not overlap at any point in the display sequence. (See Mingolla & Todd, 1984, for a detailed discussion of the computational techniques for the manipulation and display of quadric surfaces.)

Each object was rotated 90° at a constant angular velocity about a vertical axis through the center of the display

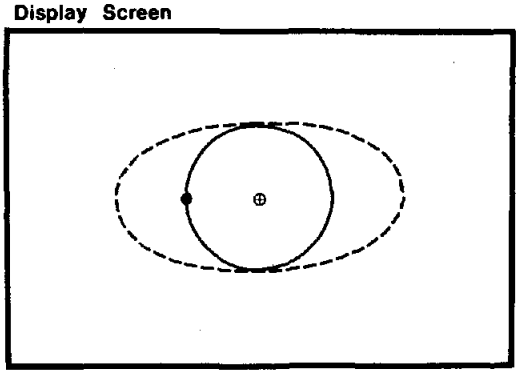

Figure 11. Two images of the same ellipsoid as viewed perpendicular to the display screen. (Note that the identifiable point, which is located on the self-occluding boundary at one moment in time, is located in the center of the image following the 90° rotation. In terms of their two-dimensional motion, the optical projections of the self-occluding contour and the identifiable point would be moving in opposite directions.)

screen. Following this 90° rotation, it would abruptly reverse its direction of motion and return to its original position. This was repeated for seven complete oscillation cycles at a rate of ½ Hz.

The moving displays were created by stepping through a sequence of 24 frames, which were generated from a simulated viewing distance of 39 m (i.e., they closely approximated a parallel projection). Each frame depicted the optical projection of the simulated object as a homogeneous blue patch against a black background. For example, Figures 12 and 13 show how the objects in Displays 3 and 4 would appear at four different points in their rotation cycles. It is important to note in these figures that the visible contours were all produced by the self-occluding boundaries of the objects, so that at each moment in time a given contour in the display would represent a different locus of points on the simulated surface. Thus, the resulting deformation did not satisfy the assumption of projective correspondence as is required by existing methods for computing structure from motion.

A wide variety of observers have viewed these displays in an informal laboratory setting. The usual experimental procedure is to obtain observers' subjective reports both before and after being informed about how the displays were generated.

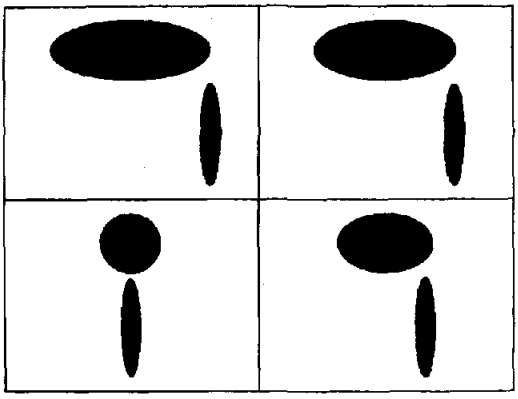

Figure 13. Four static images from the motion sequence in Display 4 of Demonstration 2. (Moving clockwise from the upper left, the images depict a nonintersecting pair of ellipsoids with rotations of 0°, 30°, 60° and 90°, respectively, from their initial orientations. When this sequence is observed in rapid succession, it produces a relatively weak kinetic depth effect.)

Results and Discussion

Let us first consider the results for the horizontal ellipsoid in Display 1. Invariably, naive observers who view this display are unable to recognize a solid object rotating in three-dimensional space. The most common impression is that of an elastic disk being stretched back and forth in the plane of the display screen. Most observers can achieve a rigid, three-dimensional interpretation of a cigar-shaped ellipsoid after hearing a description of how the display was generated, but the perceptual organization in that case is not compelling and is difficult to maintain.

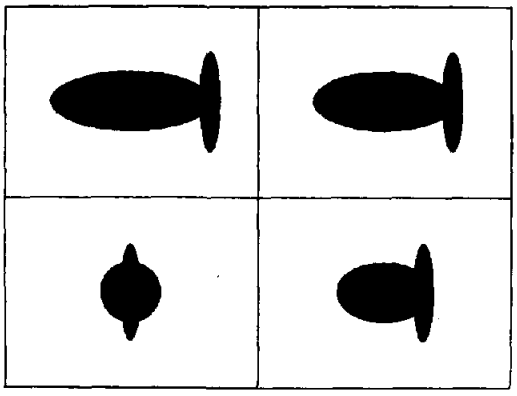

Figure 12. Four static images from the motion sequence in Display 3 of Demonstration 2. (Moving clockwise from the upper left, the images depict a surface of intersection with rotations of 0°, 30°, 60°, and 90°, respectively, from its initial orientation. When this sequence is observed in rapid succession, it produces a compelling kinetic depth effect.)

The perception of two-dimensional motion also predominates for the vertical ellipsoid in Display 2. This is not surprising. The optical projection in Display 2 moves back and forth in the image plane without any noticeable changes in size or shape. An informed observer can usually perceive a small amount of motion in depth, probably because of the sinusoidal variations in velocity (see Braunstein, 1976), but, as in the previous example, the effect is not a strong one.

Because neither of these objects presented singly produces a compelling kinetic depth effect, it would be reasonable to expect that the same would be true for the two presented in combination. That is not the case, however. The surface of intersection in Display 3 (see Figure 12) produces an immediate and dramatic impression of a solid object rotating in three-dimensional space. The details of an observer's subjective report can vary from one individual to another (e.g., it might be called a blimp, a fish, or a torpedo), but the perception of a rotating solid object is always reported, regardless of the observer's prior knowledge of how the display was constructed.

Moreover, if observers are instructed to make a conscious effort to see the display as an elastic deformation, they generally report that they are unable to do so.

Why should Display 3 produce a more compelling kinetic depth effect than Displays 1 and 2? There are two important differences between these displays that are potentially relevant to the perception of structure from motion. Note, for example, that Displays 1 and 2 depict a single object in isolation, whereas Display 3 depicts a surface of intersection that is a combination of two objects. It is possible that the yoked motion of two different objects produces a mutual constraint on the perceptual interpretation of each one (e.g., see Rock, 1983). A second difference between these displays is that the contours in Displays 1 and 2 are perfectly smooth, whereas the contours in Display 3 contain singularities where one object is temporarily occluded by another. The manner in which these singularities change over time could also provide an additional constraint for determining an object's three-dimensional structure.

One way of comparing these alternative explanations is to separate the two ellipsoids in the vertical dimension so that the yoked deformation of their contours is presented in isolation, without any changes in the pattern of contour intersections (see Figure 13). This is the purpose of Display 4. From the observers' subjective reports for this display it is clear that the resulting pattern of optical motion has a high degree of multistability. Some naive observers give an unprompted report of a pair of solid objects rotating in space, whereas others do not. For an informed observer it is generally possible to switch back and forth at will between a two- and three-dimensional perceptual organization (although the three-dimensional organization is often reported as slightly nonrigid). When asked to rank the displays in terms of the salience of the kinetic depth effect, all observers agree that Display 4 is superior to Displays 1 and 2, but that it is not nearly as compelling as the surface of intersection in Display 3.

All of this suggests that observers' judgments of these displays are influenced by at least two sources of information. The increased salience of Display 4 relative to Displays 1 and 2 indicates that the yoked movements of spatially separated contours provide some degree of information for the perception of structure from motion. Similarly, the increased salience of Display 3 relative to Display 4 indicates that a changing pattern of contour intersections can also provide perceptually useful information about an object's three-dimensional form. It is interesting to note that both of these sources of information have been implicated in other aspects of object and event perception. For example, the yoked (i.e., common) movements of spatially separated objects play an important role in current theories of perceptual organization (e.g., Restle, 1979), whereas the intersections of contours has proven to be a primary source of information for analyzing line drawings of plane-faced polyhedra (e.g., Guzman, 1968).

It is important to keep in mind that the deformations of self-occluding contours, such as those depicted in the present demonstration, could not be given a correct three-dimensional interpretation by any computational analysis that is based on an assumption of projective correspondence. Moreover, it is most unlikely that an intermediate process such as those suggested in Experiments 1 and 2 could eliminate these violations of the correspondence assumption by a simple transformation of the optical input. Note in Figures 10 and 11, for example, that at a given instant in time, a self-occluding contour may be moving in one direction, while the true optical projection of the locus of points to which it corresponds at that instant is moving in the opposite direction.

An alternative approach to the analysis of structure from motion that does not involve an assumption of projective correspondence is suggested by the work of Koenderink and van Doorn (1976). These authors have shown how a pattern of self-occluding contours within the optic array can be analyzed in terms of four basic components—spines, T-junctions, hyperbolic arcs, and elliptic arcs—that can each be related to specific aspects of topological structure for a smoothly curved surface in three-dimensional space. Sometimes, when an object is in motion, a component of its self-occluding contour may be abruptly replaced by another. Koenderink and van Doorn have shown that there are a limited number of ways in which this can happen and that they are all related in a straightforward manner to the ob-

ject's three-dimensional form. The only problem with this analysis for interpreting the results of the present demonstration is that it is qualitative in nature. It cannot, for example, distinguish between a spherical object and an elongated cigar shape, nor can it account for the perceptual significance of yoked contour deformations in Display 4. Nevertheless, despite these limitations, it is surely an important first step in the right direction.

Deformations of Scalar Fields

The movements of points and contours within a visual image are by no means the only aspects of optical structure that change over time. Under natural viewing conditions, there are likely to be extensive areas of the optic array in which there are no discernible contours at all. The pattern of image intensity within such a region will form a two-dimensional scalar field that can deform over time when an object is observed in motion (see Koenderink & van Doorn, 1980; Todd, 1985).

It is important to recognize that the global deformations of an intensity field do not satisfy the assumption of projective correspondence as is required by existing models for computing structure from motion. The projected intensity of any given surface point can be influenced by a wide variety of environmental variables, including the orientation of the surface at that point with respect to the direction of gaze and the directions of illumination (see Todd & Mingolla, 1984). Because of these many influences, the optical projection of a point on a moving surface will generally have different intensities at different moments in time. Within the overall pattern of image intensity there may be identifiable structures (e.g., the extrema produced by shadows and highlights) but, in general, these structures will not correspond over time to a fixed locus of points on an object's surface.

To avoid these violations of the correspondence assumption, some researchers (e.g., Ullman, 1979) have argued that changes in image intensity are not a direct stimulus for human motion perception. According to this view, contours must first be extracted from zero-crossings in the intensity field, and it is the movements of these contours from which motion perception is derived. This hypothesis is vitiated, however, by the following Gedanken experiment: Consider the optical projection of a rotating egg or cigar-shaped object as viewed by a monocular observer.[4] From the previous demonstration we know that the deforming self-occluding boundary of such an object in isolation does not provide sufficient information for a compelling kinetic depth effect. Unfortunately, if the object's surface is constructed with a smooth, homogeneous material that is devoid of discernible texture, then there will be no other contours within the image on which an analysis of structure from motion could be based—except perhaps those produced by shadows or highlights, which also violate the assumption of projective correspondence. In short, if the perception of structure from motion can be based only on the optical movements of points and contours, as is assumed by current models, then the optical projection of a rotating egg or cigar-shaped object should be perceived as an elastic deformation. Our experiences with solid objects under natural viewing conditions provides ample evidence that this prediction is incorrect.

One possible source of information about the movements of a solid object in three-dimensional space may be available from the deformations of its intensity field. To better understand the structure of this field and how it changes over time, it is useful to consider the three shaded images in Figure 14 together with their corresponding field diagrams in Figure 15. The upper image in Figure 14 depicts a cigar-shaped ellipsoid with its major axis oriented perpendicular to the image plane. The middle and lower images depict the same object rotated 45° and 90°, respectively, about a vertical axis through its center. All of the objects have matte (i.e., Lambertian) surfaces and are illuminated by a far-away light source near the point of observation. The intensity fields for both of these images are represented in Figure 15. The solid lines in this figure are called isointensity contours because each one represents a locus of image points that have

[4] It is important to note in this regard that self-occluding contours also pose serious problems for theories of stereopsis. William Warren and I are currently investigating the ability of human observers to cope with these problems, and we hope to report our findings in a future article.

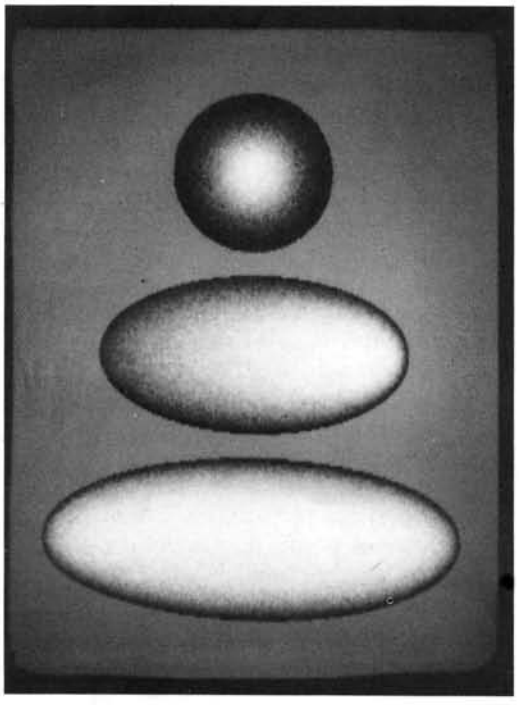

Figure 14. Three static images from the motion sequence in Demonstration 3. (The images from bottom to top depict a horizontal ellipsoid with rotations of 0°, 45°, and 90° from its initial orientation. Note that the rigid relation between these objects is not readily apparent in this static presentation. When the sequence is observed in rapid succession, however, it is immediately identified as a solid object rotating rigidly in three-dimensional space.)

identical intensities. It is important to note in these figures that as the simulated object is rotated in space, its isointensity contours within a visual image are systematically deformed. Demonstration 3 was designed to examine whether these deformations of the intensity field provide perceptually salient information about an object's three-dimensional structure.

Demonstration 3

Method

A single computer simulation was generated of a solid object rotating in three-dimensional space. The depicted object was identical to the horizontal ellipsoid in Display 1 of the previous demonstration. That is to say, it had x-y-z semiaxes of 7.8, 2.9, and 2.9 cm, respectively, in its initial orientation; it rotated back and forth in depth through a 90° angle for seven complete oscillation cycles at a rate of ½ Hz; and it was displayed at a simulated viewing distance of 39 m.

The critical difference from the previous demonstration is that the object was depicted with a smoothly varying pattern of shading. The shading was designed to simulate a matte surface, illuminated by a point-light source at the point of observation, with a 20% component of diffuse, ambient illumination. To produce this pattern of shading for any given image a different intensity value (I) was computed for each pixel in the display. These intensity values were all generated from a single equation: $I = 51 + 204 \cos \theta$, where θ was the angle between the surface normal at a depicted point and the direction of illumination at that point (see Todd & Mingolla, 1983, for a more detailed discussion of image shading).

The display sequence was composed of 24 separate images, each of which had an intensity resolution of three bits. Because of this limited intensity resolution, the full

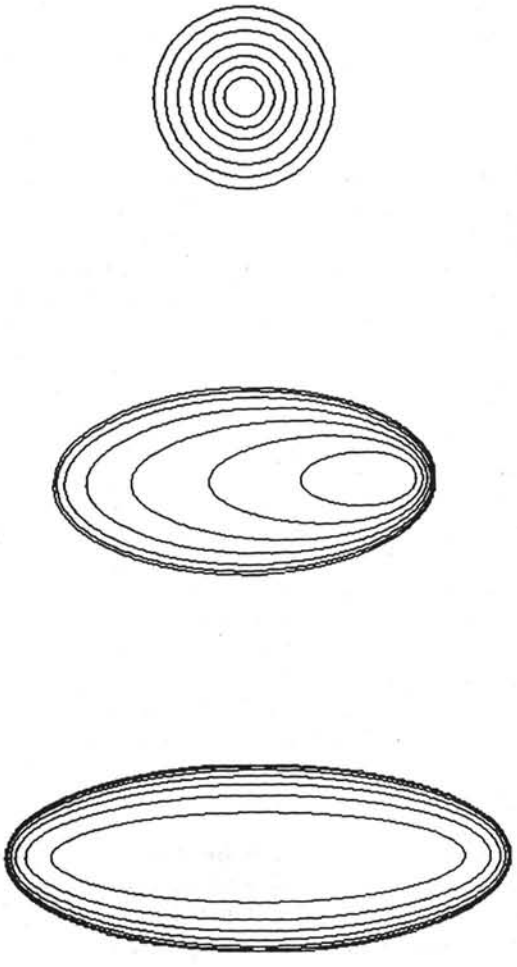

Figure 15. The intensity field diagrams for the three images in Figure 14. (The solid lines depict isointensity contours which connect points of equal intensity. Each successive contour moving outward represents a constant reduction of image intensity equal to $\frac{1}{10}$ the maximum possible value.)

range of intensity values used in the simulations could not be achieved at the level of individual pixels. To overcome this problem, a method of pseudoshading was employed. The calculated intensity value for each pixel was divided by 32 to obtain a pixel level intensity between one and seven. The remainder of this division was then compared with a randomly selected number between 0 and 31. If the remainder was smaller, then the pixel level intensity value would be reduced by one. This random component in the intensity calculations produced a gradual gradation in the overall pattern of image shading, which appeared to the observer as a slightly mottled surface (see Figure 14 for an example of the images that were generated with this procedure).

A wide variety of observers have viewed this display in an informal laboratory setting. As in the previous demonstration, the observers' subjective reports were obtained both before and after they were informed about how the display was generated.

Results and Discussion

Although the shading in this display is rather crude, the perceptual effect is, nevertheless, quite compelling. All observers report the impression of a solid object rotating rigidly in three-dimensional space, regardless of their prior knowledge of how the display was generated. Note that the perceptual effect of this display contrasts sharply with the previous demonstration in which an identical moving object was presented without shading. In that case, the object was perceived as an elastic disk being stretched back and forth in the picture plane. When shading is added to the display, however, its perceived structure is altered dramatically.

A closely related demonstration of the importance of shading for the perception of structure from motion has also been performed by William Warren (personal communication, September, 1984). Warren created a series of images of elongated ellipsoids in different orientations using detailed patterns of shading with 256 different intensity values. He then recorded these images in sequence on a video tape. When the tape was played back at high speed to eliminate jerkiness at the transition points, observers reported that it appeared quite clearly as a solid object rotating rigidly in three-dimensional space.

It is important to keep in mind that these demonstrations differ significantly from the classical kinetic depth effect reported by Wallach and O'Connell (1953) in that the depicted object appears three-dimensional even in a static presentation (e.g., see Figure 14). It might be tempting to conclude on the basis of this observation that the perception of three-dimensional form in this case is due solely to the statically available information and that the deformation of the intensity field has no effect whatsoever. If this hypothesis were correct, however, then the perceived three-dimensional form of the object should be unaffected by whether it is presented statically or in motion. The results do not confirm this prediction. Note in Figure 14, which shows three static images, that the upper object appears more like a sphere than an elongated cigar shape and that the middle object appears to be oriented parallel to the picture plane rather than at a 45° angle. This apparent regression into the display screen for objects depicted in shaded images is a general phenomenon that has been described in detail by Mingolla (1983) and Mingolla and Todd (in press), but the effect is eliminated when an object is observed in motion. Under dynamic presentation the objects depicted in Figure 14 are correctly identified as having identical shapes in different orientations. This suggests strongly that the pattern of image motion provides additional information about an object's three-dimensional structure that is not available in any of the individual images from which the motion sequence is composed.

There are at least two general methods by which continuous changes in the pattern of shading could be used to determine an object's three-dimensional structure. One approach developed by Horn & Shunck (1981) is to transform the deformations of the intensity field into a vector field of velocities for which the assumption of projective correspondence is satisfied. The output of this transformation would then be compatible with existing methods of computing structure from motion. Although this general approach is quite reasonable, the specific analysis proposed by Horn and Shunck is derived from some highly restrictive assumptions that would seldom be satisfied under natural viewing conditions. Thus, in its present form, the analysis has little value as a model of human perception.

An alternative approach to the problem that has been adopted by Koenderink and van Doorn (1980, 1982a) is to search for aspects of an object's structure that are directly spec-

ified by the deformations of its intensity field. Koenderink and van Doorn have discovered that singularities in the pattern of image intensity (i.e., maxima, minima, and saddle points) are a particularly rich source of information about the topological structure of smoothly curved surfaces in three-dimensional space. They have demonstrated, for example, that when an object is observed in motion, a saddle point in the intensity field (i.e., where an isointensity contour crosses itself) will generally trace out the optical projection of a parabolic line on the object's surface that separates regions of positive and negative Gaussian curvature (see also Koenderink & van Doorn, 1982b). Although this type of analysis has great potential as a model of human perception, it is important to keep in mind that in its present form it can only provide a qualitative description of an object's structure in terms of its local Gaussian curvature. More subtle distinctions of shape such as the difference between a sphere and an elongated ellipsoid would require some other method of analysis.

Demonstration 4

Another possible source of information about the movements of a solid object in three-dimensional space may be available from the deformations over time in its pattern of optical texture (See Gibson, 1979). To better appreciate how the texture within a visual image is organized, it is useful to conceive of the optic array as a densely structured cone of arbitrarily small solid angles, each of which is projectively related to a bounded area of an observed surface. As has recently been described by Todd and Mingolla (1984), this mapping of projected areas defines a two-dimensional scalar field that uniquely determines the global organization of optical texture. Along any isocontour in this projected area field, the optical texture elements will be homogeneously distributed, and they will all have approximately the same size and shape, except for random variations that may occur locally. In addition, all systematic changes in optical texture (i.e., the texture gradients) will be oriented in a direction that is perpendicular to these isocontours.

Under certain conditions, the projected area field that determines the pattern of optical texture for a given surface will be closely related to the field of image intensities for that surface. In fact, whenever an observed surface has a Lambertian reflectance function and is illuminated by a light source near the point of observation (e.g., when a photograph is taken with a flash camera), the two fields will be identical. Because it has already been demonstrated that the deformations of an intensity field under these conditions provide perceptually salient information about an object's three-dimensional form, it is reasonable to speculate that an equivalent effect might also be achieved by global deformations in the pattern of optical texture.

It may appear at first blush that this hypothesis contradicts the results of Experiment 1, as well as those of Pittenger (personal communication, 1979) described earlier. In both of these experiments a moving display was presented with zero correspondence between the individual elements, so that all that remained for the perception of structure from motion was the deformation over time of the global pattern of optical texture. In each case, the display failed to produce even a hint of a kinetic depth effect. It is important to keep in mind, however, that the patterns of texture employed in these experiments were composed entirely of variations in texture density, with little or no variation in the sizes and shapes of the individual elements. Because gradients of size and shape are known to be the most salient aspects of optical texture (e.g., see Cutting & Millard, 1984; Flock & Moscatelli, 1964), Demonstration 4 was designed to examine whether the deformations of these gradients can provide information about an object's three-dimensional form.

Method

A sequence of 24 separate images was generated to simulate a solid object rotating in three-dimensional space. The shape of the simulated object and its pattern of motion were identical in all respects to the horizontal ellipsoid in Demonstrations 2 and 3. The primary difference from these previous demonstrations was that the depicted object was invisible except for a set of small luminous squares that were homogeneously scattered across its surface in random orientations. In three-space, the individual square elements all had dimensions of 0.6×0.6 cm, and were distributed homogeneously over the object's surface (see Todd & Mingolla, 1984). Because of the effects of perspective, however, the elements of optical texture in the image plane exhibited systematic changes in size, shape, and density as a function

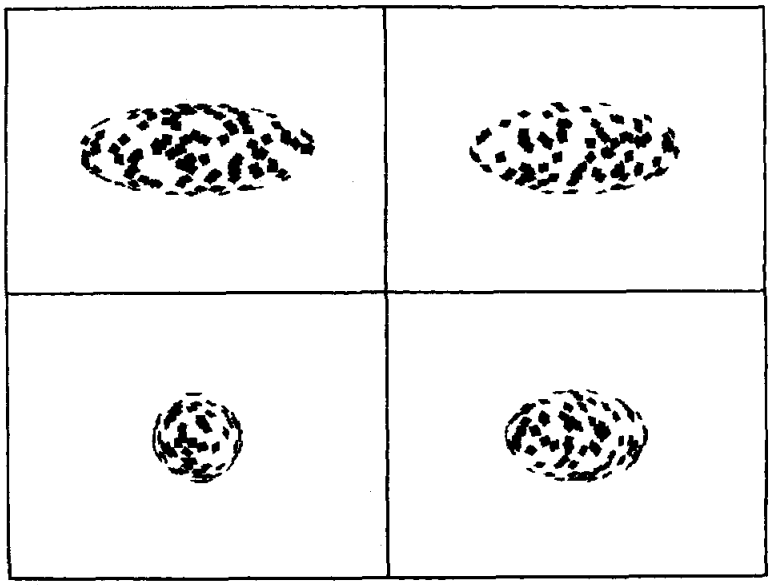

Figure 16. Four static images from the motion sequence in Demonstration 4. (Moving clockwise from the upper left, the images depict a horizontal ellipsoid with rotations of 0°, 30°, 60°, and 90° from its initial orientation. Note that the texture distributions in the different images do not correspond to one another and that the rigid relation between the objects is not readily apparent in a static presentation. When the sequence is observed in rapid succession, however, it is immediately identified as a solid object rotating rigidly in three-dimensional space.)

of surface orientation. The global organization of this optical texture was identical to the pattern of shading in Demonstration 3, and it deformed over time in exactly the same way as is represented in Figure 15. For describing patterns of texture, an isocontour in this figure would represent an idealized locus of image points along which the optical texture elements have comparable sizes, are equally foreshortened, and are uniformly distributed. To eliminate any correspondence over time of the individual elements, a different distribution of texture was generated at random for every image in the sequence (see Figure 16 for some representative images at different points in the rotation cycle). Thus, the only available information for the perception of rotation in depth was the deformation over time of the global pattern of optical texture. Each of the depicted objects in the rotation sequence contained between 75 and 100 texture elements that covered approximately 25% of its visible surface.

A wide variety of observers have viewed this display in an informal laboratory setting. As in the previous demonstrations, observers' subjective reports were obtained both before and after they were informed about how the display was generated.

Results and Discussion

The perceptual effect of this display is comparable to that of the shaded ellipsoid in Demonstration 3. That is to say, all observers report the impression of a solid object rotating in three-dimensional space, regardless of their prior knowledge of how the display was generated. For some observers, the surface appears to be scintillating with the individual texture elements appearing and disappearing at random. Others report a type of swirling motion in which the individual elements seem to move continuously over the surface in random trajectories. In either case, however, the otherwise invisible ellipsoid to which the elements are attached is clearly perceived to be rotating rigidly in three-dimensional space.

It should be noted in Figure 16 that when the different images of the motion sequence are presented individually, they do not appear to be rigidly related. In the two objects depicted on the left, for example, the one on the top looks more elongated than does the one on the bottom. Similarly, the two objects depicted on the right appear to be more closely aligned with the picture plane than is appropriate for their simulated orientations. These observations suggest that the pattern of image motion in this display provides additional information

about the depicted object's three-dimensional form that is not available in any of the individual images from which the motion sequence is composed.

The most likely source of this information is the continuous deformation in the overall pattern of image texture, which, in terms of its field structure, is identical to the continuous deformations in shading presented in Demonstration 3. Because of this close mathematical relationship between shading and texture, a field structure analysis of image intensity, such as the one proposed by Koenderink and van Doorn (1980, 1982a), should also apply equally well to patterns of image texture. This assumes, however, that the structure of the texture field can be adequately determined from the statistical sampling of texture elements that is available in any given image. There are several possible techniques by which this could be accomplished. For example, one particularly promising approach based on dynamic neural interactions is suggested by the work of Grossberg (1983).

General Discussion

During the past decade there has been a growing effort among researchers in a variety of fields to develop a computational analysis of how human observers perceive structure from motion. A fundamental problem of this research is that visual images are inherently ambiguous—that is to say, there is an infinite number of possible events in 3-space that are projectively equivalent to any given moving image on a two-dimensional display surface. Most theorists have addressed this problem by postulating constraints on the structure of the environment that limit the number of possible three-dimensional interpretations of an object to be considered. Unfortunately, most of the constraints that have been proposed to date seem to have been adopted more for their mathematical convenience than for their psychological validity. As a result, existing analyses are not easily generalized to the unrestricted patterns of stimulation that can occur under natural viewing conditions, and are, therefore, of dubious value as models of human perception (see Braunstein & Andersen, 1984; Todd, 1984).

Consider, for example, a commonly accepted assumption in the analysis of structure from motion that moving elements on the retina are the optical projections of identifiable moving points in three-dimensional space. This assumption may appear at first blush to be perfectly reasonable. Indeed, it has been so taken for granted in the literature that most investigators have not even acknowledged it as an assumption. A closer examination reveals, however, that there are many optical phenomena encountered in nature for which the assumption of projective correspondence is invalid. Thus, in light of the fact that existing computational models are unable to deal with these phenomena, the present investigation was designed to determine whether similar limitations are also exhibited by actual human observers.

The results of this research provide strong evidence that the ability of human observers to perceive structure from motion is much more general than would be reasonable to expect on the basis of current theory. Whereas existing computational models perform as advertised only within narrowly constrained boundary conditions, the available psychophysical evidence indicates that no such limitations exist for actual human observers. The present research has demonstrated, for example, that observers can experience a compelling kinetic depth effect even when the pattern of optical motion is contaminated by large amounts of visual noise (e.g., where the signal to noise ratio is less than 0.15), and that deformations of shading, texture, or self-occluding contours, which would be treated as noise by existing computational models, are analyzed by human observers as perceptually salient sources of information about an object's three-dimensional form.

It is important to keep in mind when evaluating the results of these experiments that there are several other frequently encountered violations of the correspondence assumption, such as the deformations of cast shadows or specular highlights, whose perceptual effects have yet to be examined. Taking all of these possible violations into account, one cannot help but question whether the assumption of projective correspondence is an adequate foundation for the development of perceptual

theory. Moreover, when we consider these violations in conjunction with other limitations of existing computational models, such as restrictions on viewing distance and the rigidity of an object's motion (see Braunstein & Andersen, 1984; Todd, 1984), it becomes reasonable to question whether existing models have any ecological validity whatsoever.

Metaanalysis of Visual Processing Strategies

All of this suggests that the modular analyses of visual perception that have dominated the literature in recent years (see Marr, 1982) will have to be modified if they are to account for the high level of generality exhibited by human observers. The main problem with this modular approach is that the individual modules proposed thus far are so limited that they produce erroneous outputs over a significant range of environmental conditions. Thus, if a modular theory is to be salvaged at all, then it must include some type of mechanism for identifying the particular modules that are appropriate in any given situation. Because of the significance of this issue for perceptual theory in general, it is worthwhile to consider briefly some possible strategies by which it could be addressed.

Executive processes. One way of ensuring that a specialized processing module does not impair perceptual performance when viewing conditions fail to satisfy its underlying assumptions is to continually monitor whether the output of that module is consistent with an observer's expectations based on general knowledge. This monitoring function would presumably be performed by some sort of executive process, which would be responsible for the final determination of an object's three-dimensional form. Although this strategy may at first seem quite promising, there are a number of severe difficulties in its actual implementation. Psychologists have traditionally invoked executive processes to explain all manner of perceptual phenomena (e.g., see Rock, 1983), but the precise details of how these processes function are seldom specified. How, for example, would the executive process sort through the vast quantities of information in human memory to find just the right piece of knowledge that is appropriate in any given situation, and what are the specific criteria by which this knowledge would be used to override more specialized processing modules in the analysis of visual information? The difficulty of these problems makes a monitoring strategy seem much less attractive. Moreover, there is considerable evidence in the literature that higher order knowledge often has surprisingly little influence on the processes of human perception. The visual illusions are a compelling case in point. For example, although we may be fully aware that an observed object is in reality a rotating trapezoid, it is still perceived as an elastically deforming rectangle (Ames, 1951). Such findings suggest the operations of executive processes may be much less significant to human vision than has traditionally been assumed.

Context-dependent processes. Another possible strategy for overcoming the limitations of a specialized processing module is to make its operations dependent on other perceptual analyses, which are designed to determine whether current viewing conditions satisfy its underlying assumptions. For example, if an analysis of structure from motion can be based on an assumption of projective correspondence (or object rigidity), as suggested by current theory, then perhaps there is some form of information by which projective correspondence (or object rigidity) is visually specified. Some of the computational models reported in the literature do indeed contain auxilliary analyses for testing their underlying assumptions (e.g., see Lee, 1974; Todd, 1982), but there are some potential difficulties with this approach that need to be considered. Notice, for example, the danger of an indefinite regress. If the assumptions of one analysis must be tested by another, then the assumptions of the second analysis would have to be tested as well, and so on. Another important difficulty with this type of context-dependent processing is that it does nothing to achieve the high level of generality that is so characteristic of human perception. Although there may be visual information to verify the underlying assumptions of a specialized processing module, a more general purpose device would still seem to be necessary for those situations where its assumptions are violated. If that device were sufficiently general, moreover, then the existence

of a specialized processing module would be functionally superfluous (see Todd, 1984, for further discussion of this issue).

Competitive/cooperative processes. A third possible strategy for overcoming the limitations of a specialized processing module is to have it interact with other related processes. This approach assumes that objects and events in a natural environment can be multiply specified by many different sources of information, each of which is detected by a specialized processing module with its own individual limitations. In any given situation, we would expect to obtain erroneous outputs from some of these modules because of inappropriate viewing conditions, but it would be most unlikely for two or more of them to fail in exactly the same way. Thus, if the different modules could be designed to excite one another when their outputs are compatible and to inhibit one another when their outputs are incompatible, then the inappropriate modules would be dynamically suppressed, and the system would eventually converge on a correct interpretation of the available information (a more detailed discussion of dynamic processes in visual perception can be found in Grossberg, 1983, and Grossberg & Mingolla, 1985). A particularly desirable property of this general strategy is that the individual processing modules could be relatively crude yet still contribute positively to the overall function of the entire system. This would allow for heuristic processes as described by Braunstein (1976)—see also Todd and Warren (1982)—and would be highly conducive to the process of evolution (see Braunstein, 1983).

It is important to keep in mind that the three processing strategies described above are not mutually exclusive and that they do not necessarily exhaust all of the possible strategies that are potentially available. The fact remains, however, that something of this sort will be required if existing computational analyses are to be taken seriously as models of human perception. Until now, perceptual theorists have been behaving much like the proverbial drunk who has lost his keys in the shadow of a building but searches for them under a street lamp because it is easier to "see" there. The street lamp in this story is analogous to the dubious and highly restrictive assumptions on which existing computational models are so precariously based. If, like the drunk, we are to find the keys to the processes of human vision, then we will have to face up to the entire range of optical phenomena that are encountered in a natural environment, including the effects of visual noise, self-occluding contours, and patterns of shading.

Rediscovery of the Optic Array

Although most of the discussion in the present article has been presented from the perspective of computational theory, it is only fair to acknowledge that a similar set of conclusions was arrived at by Gibson (e.g., 1961, 1966, 1979) over 20 years ago—long before the current spate of computational models began to appear in the literature. Indeed, Gibson's realization of the enormous variety of structure in ambient light was a primary determinant for developing his concept of the optic array. Consider, for example, the following passage from *The Ecological Approach to Visual Perception* (1979), in which he considers how different aspects of optical structure can change over time:

> Can these disturbances of structure be treated mathematically? They surely cannot all be treated with the same mathematical method, for some of them do not conform to the assumptions of the theory of sets. Some of the above changes do not preserve a one-to-one mapping of units over time, inasmuch as the array gains or loses units in time. Accretion or deletion of texture during occlusion is one such case. Foreshortening or compression of texture preserves one-to-one mapping only until it reaches its limit, after which texture is lost. The emergence of new texture with rupturing of a surface, the nullification of texture with dissipation of a surface, and the substitution of new texture for old are still other cases of the failure of one-to-one mapping, or projective correspondence. In all of these cases it is not the fact that each unit of the ambient array at one time goes into a corresponding unit of the array at a later time. The case of an optic array that undergoes "flashing" or scintillation of its units is another example, and so is what I called fluctuation in connection with changing light and shade. (p. 108)

It should be clear from this passage that Gibson would not have been surprised by the results of the present experiments and that he anticipated the difficulties such results would pose for a computational theory. Gibson conceived of these many varieties of optical structure and their changes over time not as an impediment to perception as suggested by more recent theorists, but as potential sources of information about objects and events in a natural

environment. The research described in the present article provides strong evidence that his insightful conception of the nature of visual information was fundamentally sound.

To develop Gibson's insights about the optic array into a complete theory of visual perception, future theorists will have to provide a formally precise account of how specific disturbances of optical structure relate to the environment and how they are exploited as sources of information by the human visual system. For example, one promising direction for theoretical development suggested by the present research is to analyze the deforming field structures of optical properties, such as shading or texture, that vary continuously in visual space. As was described in Demonstrations 3 and 4, the geometric structures of these fields can be closely related even though they are defined over different optical properties, thus providing a potentially useful source of converging information. The mathematical analysis of these optical field structures has already been initiated by Koenderink and van Doorn (1975, 1976, 1977, 1980, 1982a), with considerable success, but there is much that remains for future research.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatio-temporal energy models for the perception of motion. *Journal of the Optical Society of America. 2,* 284–299.

Ames, A. (1951). Visual perception and the rotating trapezoidal window. *Psychological Monographs, 67*(7, Whole No. 324).

Braunstein, M. L. (1976). *Depth perception through motion.* New York: Academic Press.

Braunstein, M. (1983). Contrasts between human and machine vision: Should technology recapitulate phylogeny? In J. Beck, & A. Rosenfeld (Eds.), *Human and machine vision* (pp. 85–96). New York: Academic Press.

Braunstein, M. L., & Andersen, G. J. (1984). Shape and depth perception from parallel projections of three-dimensional motion. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 749–760.

Cutting, J. E., & Millard, R. T. (1984). Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General, 113,* 198–216.

Doner, J., Lappin, J. S., & Perfetto, G. (1984). Detection of three-dimensional structure in moving optical patterns. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 1–11.

Doorn, A. J. van, & Koenderink, J. J. (1982a). Spatial properties of the visual detectability of moving spatial white noise. *Experimental Brain Research, 45,* 189–195.

Doorn, A. J. van, & Koenderink, J. J. (1982b). Temporal properties of the visual detectability of moving spatial white noise. *Experimental Brain Research, 45,* 179–188.

Doorn, A. J. van, & Koenderink, J. J. (1982c). Visibility of movement gradients. *Biological Cybernetics, 44,* 167–175.

Doorn, A. J. van, & Koenderink, J. J. (1983). Detectability of velocity-gradients in moving random-dot patterns. *Vision Research, 23,* 799–804.

Flock, H. R. (1964). Some conditions sufficient for accurate monocular perceptions of moving surface slants. *Journal of Experimental Psychology, 67,* 560–572.

Flock, H. R., & Moscatelli, A. (1964). Variables of surface texture and accuracy of space perceptions. *Perceptual and Motor Skills, 19,* 327–334.

Gibson, J. J. (1961). Ecological optics. *Vision Research, 1,* 253–262.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Gibson, J. J., & Gibson, E. J. (1957). Continuous perspective transformations and the perception of rigid motion. *Journal of Experimental Psychology, 54,* 129–138.

Gibson, E. J., Gibson, J. J., Smith, O. W., & Flock, H. (1959). Motion parallax as a determinant of perceived depth. *Journal of Experimental Psychology, 58,* 40–51.

Grossberg, S. (1983). The quantized geometry of visual space: The coherent computation of depth, form, and lightness. *The Behavioral and Brain Sciences, 6,* 625–692.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of form perception: Illusory figures and neon color spreading. *Psychological Review. 92,* 173–211.

Guzman, A. (1968). *Computer recognition of three-dimensional objects in a visual scene.* Doctoral dissertation (MAC-TR-59, Project MAC), Massachusetts Institute of Technology, Cambridge, MA.

Hildreth, E. C., (1983). *The measurement of visual motion.* Cambridge, MA: MIT Press.

Horn, B. K. P., & Shunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17,* 185–203.

Koenderink, J. J., & van Doorn, A. J. (1975). Invariant properties of the motion parallax field due to the motion of rigid bodies relative to the observer. *Optica Acta, 22,* 773–791.

Koenderink, J. J., & van Doorn, A. J. (1976). the singularities of the visual mapping. *Biological Cybernetics, 24,* 51–59.

Koenderink, J. J., & van Doorn, A. J. (1977). How an ambulant observer can construct a model of the environment from the geometrical structure of the visual flow. In G. Hauske & F. Butenandt (Eds.), *Kybernetik* (pp. 224–247). Munich: Oldenberg.

Koenderink, J. J. & van Doorn, A. J. (1980). Photometric invariants related to solid shape. *Optica Acta, 27,* 981–996.

Koenderink, J. J., & van Doorn, A. J. (1982a). Perception of solid shape and spatial lay-out through photometric invariants. In R. Trappl (Ed.), *Cybernetics and systems research,* (pp. 943–948). Amsterdam: North-Holland.

Koenderink, J. J., & van Doorn, A. J. (1982b). The shape of smooth objects and the way contours end. *Perception, 11,* 129–137.

Lappin, J. S., Doner, J. F., & Kottas, B. (1980). Minimal conditions for the visual detection of structure and motion in three dimensions. *Science, 209,* 717–719.

Lappin, J. S., & Fuqua, M. A. (1983). Accurate visual measurement of three-dimensional moving patterns. *Science, 221,* 480–482.

Lee, D. N. (1974). Visual information during locomotion. In R. B. McLeod & H. Pick (Eds.), *Perception: Essays in honor of James Gibson* (pp. 250–267). Ithaca, NY: Cornell University Press.

Longuet-Higgens, H. C., & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London, 208,* 385–397.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Mingolla, E. (1983). *Perception of shape and illuminant direction from shading.* Unpublished doctoral dissertation, University of Connecticut.

Mingolla, E., & Todd, J. T. (1984). Computational techniques for the graphic simulation of quadric surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 740–745.

Mingolla, E., & Todd, J. T. (in press). Perception of solid shape from shading. *Biological Cybernetics.*

Petersik, J. T. (1979). Three-dimensional object constancy: Coherence of a simulated rotating sphere in noise. *Perception & Psychophysics, 25,* 328–337.

Restle, F. (1979). Coding theory and the perception of motion configurations. *Psychological Review, 86,* 1–24.

Rock, I. (1983). *The logic of perception.* Cambridge, MA: MIT Press.

Todd, J. T. (1981). Visual information about moving objects. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 795–810.

Todd, J. T. (1982). Visual information about rigid and nonrigid motion: A geometric analysis. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 238–252.

Todd, J. T. (1984). The perception of three-dimensional structure from rigid and nonrigid motion. *Perception & Psychophysics, 36,* 97–103.

Todd, J. T. (1985). Formal theories of visual information. In W. H. Warren & R. E. Shaw (Eds.), *Persistence and change: Proceedings from the first international conference on event perception* (pp. 87–102). Hillsdale, NJ: Erlbaum.

Todd, J. T., & Mingolla, E. (1983). Perception of surface curvature and direction of illumination from patterns of shading. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 583–595.

Todd, J. T., & Mingolla, E. (1984). The simulation of curved surfaces from patterns of optical texture. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 734–739.

Todd, J. T., & Warren, W. (1982). Visual perception of relative mass in dynamic events. *Perception, 11,* 325–335.

Ullman, S. (1979). *The interpretation of visual motion.* Cambridge, MA: MIT Press.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45,* 205–217.

Received May 13, 1985
Revision received July 26, 1985 ■