Journal of Experimental Psychology:
Human Perception and Performance
1989, Vol. 15, No. 4, 826-840

0096-1523/89/$00.75

Kinetic Depth Effect and Identification of Shape

George Sperling and Michael S. Landy
New York University

Barbara A. Dosher
Columbia University

Mark E. Perkins
New York University

We introduce an objective shape-identification task for measuring the kinetic depth effect (KDE). A rigidly rotating surface consisting of hills and valleys on an otherwise flat ground was defined by 300 randomly positioned dots. On each trial, 1 of 53 shapes was presented; the observer's task was to identify the shape and its overall direction of rotation. Identification accuracy was an objective measure, with a low guessing base rate, of the observer's perceptual ability to extract 3D structure from 2D motion via KDE. (1) Objective accuracy data were consistent with previously obtained subjective rating judgments of depth and coherence. (2) Along with motion cues, rotating real 3D dot-defined shapes inevitably produced a cue of changing dot density. By shortening dot lifetimes to control dot density, we showed that changing density was neither necessary nor sufficient to account for accuracy; motion alone sufficed. (3) Our shape task was solvable with motion cues from the 6 most relevant locations. We extracted the dots from these locations and used them in a simplified 2D direction-labeling motion task with 6 perceptually flat flow fields. Subjects' performance in the 2D and 3D tasks was equivalent, indicating that the information processing capacity of KDE is not unique. (4) Our proposed structure-from-motion algorithm for the shape task first finds relative minima and maxima of local velocity and then assigns 3D depths proportional to velocity.

In 1953, Wallach and O'Connell described a depth percept derived from motion cues that they called the *kinetic depth effect* (KDE). Since that time, there has been a great deal of research on the KDE, examining the effects of stimulus parameters such as dot numerosity in multidot displays (Braunstein, 1962; Green, 1961), frame timing (Petersik, 1980), occlusion (Andersen & Braunstein, 1983; Proffitt, Bertenthal, & Roberts, 1984), the detection of nonrigidity in the three-dimensional form most consistent with the stimulus (Todd, 1982), and veridicality of the percept (Todd, 1984, 1985).

Since 1979, there have been numerous attempts at modeling how observers and machines could derive three-dimensional (3D) structure from two-dimensional (2D) motion cues. Ullman (1979) referred to this computational task as the *structure-from-motion* problem. Ironically, Ullman's model and most ensuing ones do not explicitly use motion cues. These models are essentially geometry theorems concerning the minimal number of points and views needed to specify the shape under various simplifying constraints such as assumed object rigidity and assumed parallel perspective (Bennett & Hoffman, 1985; Hoffman & Bennett, 1985; Hoffman & Flinchbaugh, 1982; Ullman, 1979; Webb & Aggarwal, 1981). From the geometric models, iterative models have been developed that use newly arrived position data, not to derive the true structure, but to improve the current 3D representation in the sense of maximizing its rigidity (Landy, 1987; Ullman, 1984). Only a few models actually use point velocity (i.e., an optic flow field) in addition to point position (e.g., Clocksin, 1980; Koenderink & van Doorn, 1986; Longuet-Higgins & Prazdny, 1980), and one model also uses point acceleration (Hoffman, 1982).

It has been difficult to relate models of the KDE to the results of psychological studies. An important component of the problem has been the difficulty of finding an appropriate experimental paradigm. Many KDE experiments have used subjective ratings of "depth" or "rigidity" or "coherence" as the responses (see Dosher, Landy, & Sperling, 1989, for a review). Relating subjective responses to a process model of KDE is problematic. Typically, a structure-from-motion model yields a shape specification. To link the derived shape to subjective judgments, and thereby to experimental results, a decision-making apparatus to predict judgments is needed, and this may be quite complex.

Objective Measurements of KDE: Problems

Because the ability to derive structure from motion presumably evolved to solve an objective environmental problem, a better approach to studying KDE is to measure the accuracy of the KDE in an objective fashion. Does the observer perceive the correct shape in a display? The correct depths? The correct depth order? The correct curvature? Some of the studies cited earlier attempted to answer such questions by using objective response criteria (e.g., percentage correct in a one- or two-interval forced-choice task). Unfortunately, in almost every case, subjects can achieve good performance on the task by

The work described in this article was supported by The Office of Naval Research, Grant N00014-85-K-0077, and by the U.S. Air Force Life Sciences Directorate, Visual Information Processing Program Grants 85-0364 and 88-0140.

Correspondence concerning this article should be addressed to George Sperling, Psychology Department, New York University, 6 Washington Place, Room 980, New York, New York 10003.

neglecting perceived depth and consciously or unconsciously formulating their responses on the basis of other cues. In these cases, there is a simple non-KDE cue sufficient to make the judgment accurately. Although the subject may not consciously be using these artifactual cues to make correct judgments, we cannot be sure of the basis of the response until the artifactual cues have been eliminated or rendered useless (e.g., through irrelevant variation).

Let us consider some examples. Lappin, Doner, and Kottas (1980) presented subjects with a two-frame representation of dots randomly positioned on the surface of an opaque rotating sphere displayed by polar projection. On the second frame, a small percentage of the dots were deleted and replaced with new random dots. Subjects were required to determine which of two such two-frame displays had a higher signal-to-noise ratio (in terms of dot correspondences). Lappin et al. (1980) interpreted their results in terms of the "minimal conditions for the visual detection of structure and motion in three dimensions" (p. 717), which is the title of their article. Indeed, the signal dots represent two frames of a rigid rotating sphere. But, subjects do not need to correctly perceive a 3D sphere in order to make a correct response. There was no analysis offered of how far a 3D perception could diverge from spherical and still yield the observed accuracy of response. Alternatively, subjects might base their responses on perceived 2D flow fields, judging the percentage of dots in the first frame that have corresponding dots in the second frame. This 2D judgment need not use the entire motion flow field. For example, the 5.6° 3D motion of the sphere corresponds to a small, essentially linear translation in the center of the field. Discriminating signal-to-noise ratios in translations is related to Braddick's (1974) "dmax" procedures for discriminating perceived linear motion; it does not necessarily have anything to do with KDE. Thus, although Lappin et al. used response accuracy as their dependent variable, the subject's ability to estimate a signal-to-noise ratio may have been artifactual and certainly is not easily converted into an estimate of the accuracy of KDE.

Petersik (1979, 1980) represented rotating spheres by surface elements that were dots or small vectors. In both studies, the spheres were displayed with polar projection, and subjects were required to discriminate clockwise from counterclockwise rotation. A possible artifact here is that the motion of a single stimulus element provides sufficient information to respond correctly. That is, under polar perspective, stimulus points follow elliptical paths in the image plane. To determine rotation direction, the subject needs only determine the 2D rotation direction of a single point (assuming knowledge of the vertical position of the point with respect to eye level). Petersik made the task more difficult by adding noise to some dot paths, by varying the slant of vector elements from frame to frame, or by varying the numerosity. However, none of these manipulations prevents the subject from using a purely 2D, non-KDE strategy. Indeed, Braunstein (1977) had previously examined precisely this point. Braunstein demonstrated that only the vertical component of the polar perspective transformation was used by subjects for a depth-order judgment, and that this component was sufficient.

Andersen and Braunstein (1983) also used discrimination of rotation direction to evaluate KDE. Their displays represented clumps of dots on the surface of a sphere. A clump was construed as being bounded by an invisible pentagon, whose presence was made known by the fact that, when it lay on the front surface of the sphere, it occluded dots that lay behind it on the rear surface. These spheres were displayed by parallel perspective, and the cue to depth order (front, rear) was provided by occlusion. Again, although the dependent variable was response accuracy, a subject did not need to perceive a 3D object to determine the direction of rotation—the subject needed only to determine the movement direction of the continuously visible clumps.

In several studies, simple relative velocity cues are all that the subject needs to perform the KDE task. Braunstein and Andersen (1981) displayed a multidot representation of a dihedral edge that moved horizontally. The dots were displayed using polar projection, so that horizontal point velocities were inversely proportional to depth. Thus, the display contained a velocity gradient that either increased or decreased from the midline of the display to the upper and lower edges of the display. Subjects judged whether a given display represented a convex or concave edge. In this task, comparing the relative velocity of points in the center and at the top edge of the display is all that is necessary to perform accurately (the location with the greater velocity is judged "forward").

In experiments by Todd, subjects determined which of five curvatures (Todd, 1984) or slants (Todd, 1985) were depicted in a multidot display. Again, Todd described the task in terms of the perceived 3D object, but accurate performance is possible by comparing the relative velocities of points in just two areas of the display.

In all the studies just cited, the subject could perform the required KDE task by using a minimal artifactual cue. One possible solution to the problem of subjects learning to use artifactual cues is to withhold feedback. The assumption is that, without feedback, the subject will use only perceived 3D shape. This approach has been used extensively by Todd (1982, 1984, 1985). Unfortunately, withholding feedback does not mean that the subject cannot use an alternative perceptual or decision strategy to supplement judgments of perceived KDE depth. One strategy that subjects often adopt without feedback is to adjust their responses so as to respond equally (or nearly equally) often with each of the possible responses. For example, Todd's (1984) procedure is vulnerable to this artifact of strategy. He used surface dots to represent cylinders with five different curvatures. On a given trial, subjects judged which of the five curvatures was presented. As an alternative to perceiving KDE depth, a subject could judge the apparent velocity of dots in the center of the display and use the knowledge of the velocities displayed on previous trials to choose a curvature category. Indeed, subjects are extremely good at estimating the mean velocity and variations from it in a sequence of displays (McKee, Silverman, & Nakayama, 1986). Although the subjects' use of a trivial strategy that estimates just a single velocity per trial may not explain the entirety of Todd's results, it predicts the nearly veridical

character of subject responses and thereby could account for most of the data.

Objective Measurement of KDE: Proposed Solution

The KDE is a perceptual phenomenon that allows subjects to perceive the relative depth of different positions in visual space and hence to infer the shapes of objects in the environment. In all of the experiments we have discussed, the shapes presented were very simple (spheres, cylinders, and planes), and hence simple response strategies would have been effective. None of the experiments discussed above requires the subject to use a perceived 3D shape in order to perform accurately. In all of the studies we reviewed, subjects had the opportunity to use artifactual cues. None of these experiments presented shapes with complexity approaching that seen in the real world in which the ability to compute structure from motion evolved.

In this article, we describe a new method for investigating KDE. Our aim is to provide, instead of the demonstration of KDE by means of perceptual reports (what subjects say they see), a test of perceptual abilities (what complex shape properties subjects can extract from visual flow fields). The task is shape identification, in which on each trial, one of a large lexicon of shapes is presented. Each shape consists of a flat ground with zero, one, or two bumps or depressions. The bumps and depressions vary in position, 2D extent, and orientation. Because of the way the lexicon of shapes is constructed, good performance in the shape identification task requires simultaneous local computation of velocity in many positions of the display and global coordination of the local information.

Experiment 1: Dot Numerosity and Bump Heights

To demonstrate the shape identification method and to investigate its limits, we replicated and extended one of the classic findings in multidot KDE: the dependence of quality ratings (usually combined coherence and rigidity, or "goodness") on dot numerosity (Braunstein, 1962; Dosher et al., 1989; Green, 1961; Landy, Dosher, & Sperling, 1985). Quality of KDE generally has been found to increase with dot numerosity. We investigated the effects of dot numerosity and depth extent on the effectiveness with which subjects used the KDE to identify the target shape from among its many close competitors.

Method

Subjects. Three subjects were used in the study. Two were authors of this article, and the third was a graduate student naive to the purposes of the experiment. Two subjects had normal or corrected-to-normal vision; one subject (CFS) had vision correctable only to 20:40.

Displays. The shapes used in the experiment were 3D surfaces consisting of zero, one, or two bumps or concavities on an otherwise flat ground. Here we use the term *shape* to indicate the positions of these bumps and concavities on the flat ground, irrespective of other stimulus parameters that were varied, including bump height, number of dots used to represent the shape, and rotation direction. The shapes were constructed as follows (see Figure 1A). Within a square area with sides of length s, a circle with diameter $0.9s$ was centered. All depth values outside the circle were set to zero (i.e., in the object base plane, which in the initial display was the same as the image plane). For each of three positions inside the circle (located at the vertices of an equilateral triangle), the depth was specified as either $+h$ (a distance h in front of the object base plane, closer to the observer), 0 (in the object base plane), or $-h$ (behind the object base plane). A smooth spline was constructed, using a standard cubic spline algorithm, which passed through the flat surround and the vertices of the triangle. For a given set of vertices, 27 shapes were constructed in this way (see Figure 1B for some examples).

Two different sets of vertices were used to generate shapes. These were either at the corners of a triangle pointing up (designated u) or of a triangle pointing down (designated d). Shapes were denoted by indicating the trio of positions (u or d), and then specifying for each position (in the order shown in Figure 1A) whether that position was in front of the object base plane (+), in the plane (0), or behind it (−). For example, the shape denoted by $u+-0$ consists of a bump in the upper central area of the display, a depression in the lower left of the shape, and a flat area in the lower right of the shape (see Figure 1B). Note that $u000$ and $d000$ both designate the same shape: a flat square. Fifty-three distinct shapes can be generated in this manner.

Displays were generated for all combinations of the 53 shapes, three dot numerosities, and three bump heights. For the flat shape (denoted $u000$ or $d000$), varying bump height has no effect, and so there are only three flat shape display types (corresponding to the three numerosities). For all other shapes there are nine display types. This results in 471 display types. For most display types, a single instantiation was generated (choosing a set of random dots and forming a display after rotation and projection). For each of the display types for the flat shape, six instantiations were made. Thus, there were 486 different displays. Bump height, h, was $0.5s$, $0.15s$, or $0.05s$, where s is the length of a side of the square ground. The 3D perspective drawings of the shapes in Figure 1B are for the largest bump heights. Dot numerosities were 20, 80, and 320. The bump height and dot numerosity manipulations are illustrated in Figures 1C and 1D, respectively.

Multidot displays of these shapes were generated by choosing a random sample of positions on each surface, rotating the resulting set of points about a fixed vertical axis, and projecting them onto an image plane via parallel projection. The 3D motion was a single cycle of a sinusoidal rotation about a fixed vertical axis through the center of the object base plane, with amplitude of 25° and period of 30 frames. More specifically, the angle at which the base plane was oriented with respect to the image plane was $\theta(m) = \pm 25 \sin(2\pi m/30)$ degrees, where m is the frame number within the 30 frame display.

Two rotation directions were used, indicated as l and r, corresponding to whether the left or right edge of the display came forward initially. Equivalently, this described the side of the observer to which the shape "faced" in the second half of the rotation (which was usually an easier way to code the response). For an l rotation (see Figure 1E), the object initially appeared face-forward. It was then rotated so that the front moved to the right until the object had rotated 25°. Then it reversed direction and rotated to the left until it was 25° to the left of its initial orientation. Finally, it again reversed direction and rotated until the ground plane was again perpendicular to the line of sight. A full description of a display by a subject included the indication of the set of vertices (u or d), the 3D depths at these vertices (+,−,0), and the direction of rotation (l or r), for example, $u+-0l$.

Because of the parallel projection, simultaneous reversal of depth signs and of rotation direction yields precisely the same physical image sequence. The 486 displays described earlier were all generated

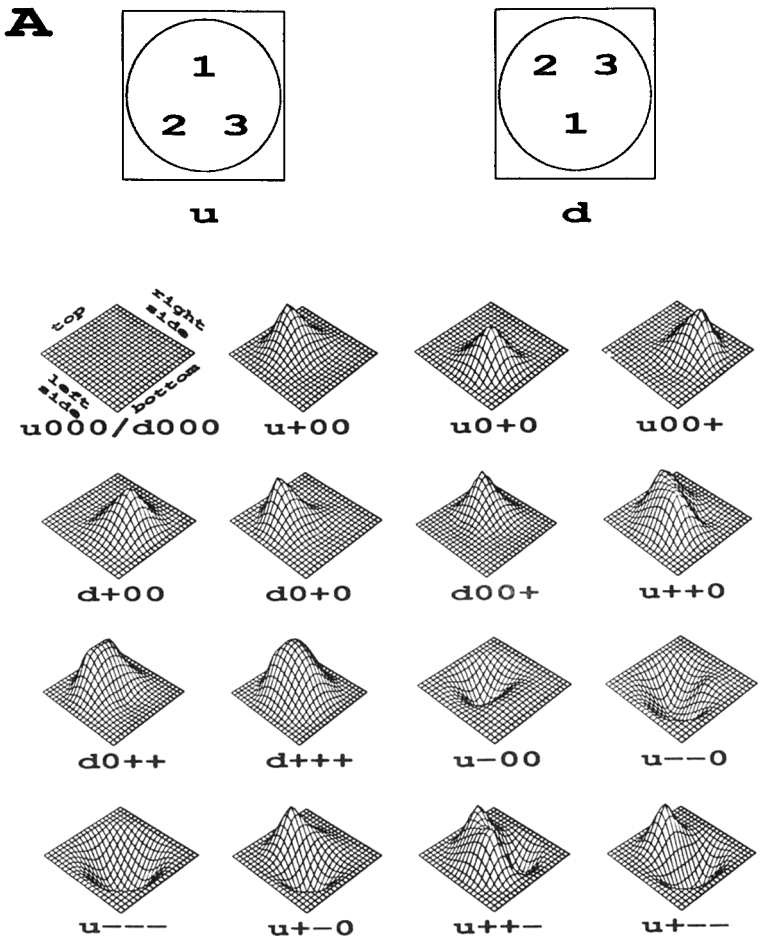

Figure 1. Stimulus shapes, rotations, and their designations. (Shapes were constructed by smoothly splining a flat ground and three points that were either toward the observer [plus sign], in the flat ground [zero], or away from the observer [minus sign].) A: These three points were at the corners of one of two possible equilateral triangles, for which the odd point is up [*u*] or the odd point is down [*d*]. In the experiment, subjects were required to name the shape and rotation direction perceived. The numbers specify the order in which the depth signs of the three points are to be reported. B: The various combinations result in a lexicon of 53 shapes; typical examples are illustrated here as perspective plots. The orientation of these plots relative to the viewing direction is indicated on the first example.

(Figure continues)

with the *l* rotation, but each can equally well be described as an *r* rotation of the sign-reversed shape. There are 108 ways to designate a display by combining an up or down shape-type with a bump, depression, or flat surface at three different locations with a left or right initial direction of motion; that is, $\{d, u\} \times \{+, -, 0\}^3 \times \{l, r\}$. For most shapes, there are two equally valid ways to describe the display. For example, *u*+−0*l* and *u*−+0*r* describe the same display. The flat shape is denoted equally accurately as *u*000*l*, *u*000*r*, *d*000*l*, and *d*000*r*. Given the four instantiations of the flat shape, chance performance depends on subject strategy. Repeated responses of *u*000*l* (and its equivalents) yields a guaranteed performance of 18 in 486 correct (or 2 in 54). Random guessing yields an expected performance of just over 1 in 54 correct. Subjects did not designate bump height in their responses. Except in the case of the flat stimuli, bump height was obvious.

After sampling, rotation, and projection, any given frame of the display consisted of *n* points in the image plane. These points were displayed as bright dots on a dark background. The square image extent of the displays projected to a 182 × 182 pixel area subtending 4° of visual angle. The displays were not windowed in any way, so the edges of the display oscillated in and out with the rotation. With the 25° wiggle, at the instants when rotation reverses, the display has shrunk to 90% of its initial horizontal extent.

Displays were presented on a background that was uniformly dark (approximately 0.001 cd/m^2). Dots were single pixels of approximately 65 μcd and were viewed from a distance of 1.6 m. A trial sequence consisted of a cue/fixation spot presented for 1 s, a 1-s blank interval, and the 2-s stimulus sequence. The stimulus sequence was followed by a blank screen, the luminance of which was the same as the background of the stimulus. The display was run at 60 Hz noninterlaced. Each display frame was repeated four times, for an effective rate of 15 new frames per second. The duration of each 30-frame display was 2 s.

Apparatus. Stimuli were computed in advance of the session and stored on disk. The stimuli were processed for display by an Adage RDS-3000 image display system and were displayed on a Conrac

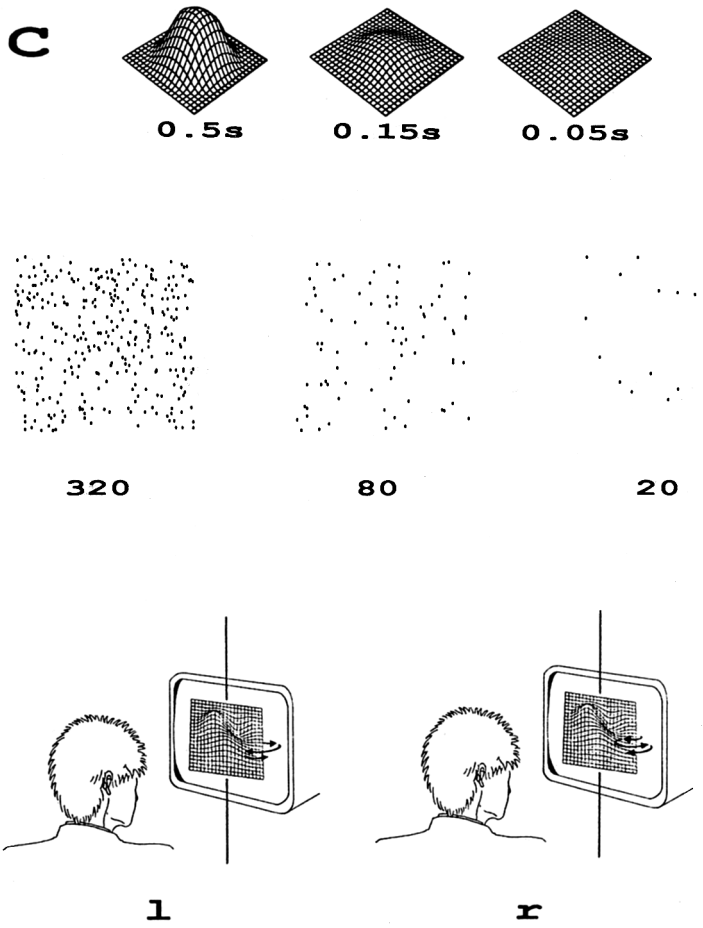

Figure 1 (continued). C: Three bump heights were used: 0.5*s*, 0.15*s*, and 0.05*s*, where *s* is the length of a side of the square base of the shape. The shape depicted here is *u*+++. D: Three dot numerosities were used: 20, 80, and 320. Pictured are the first frames of a representative display in each numerosity condition. E: Two rigid rotation motions were simulated. Both were sinusoidal rotations about a vertical axis through the center of the object ground. The object either first rotated to face the subject's right, then to the subject's left, then returned face-forward [*l*], or in the opposite direction [*r*].

7211C19 RGB color monitor. The stimuli appeared as white dots on a black background.

Viewing conditions. Stimuli were viewed monocularly (with the dominant eye) through a black-cloth viewing tunnel. In order to minimize absolute distance cues, a circular aperture slightly larger than the square display area restricted the field of view. Stimuli were viewed from a distance of 1.6 m. After each stimulus presentation, the subject typed a response on a computer terminal. Room illumination was dim. (Illuminance was approximately 8 cd/m^2.)

Procedure. Subjects were shown perspective drawings of the shapes (as in Figure 1B) and were instructed as to how they were constructed and named. They were told that they would be shown multidot versions of these shapes and would be required to name the shape displayed and its rotation direction as accurately as possible. They were told to use any method they chose to remember and apply the shape and rotation designations.

Each of the 486 displays was viewed once by each subject. The displays were presented in a mixed-list design in four sessions of 45 min each. After each response, the possible correct responses were listed as feedback. For each stimulus, there were always two responses that were scored as correct (given perceptual reversals). For the flat stimuli, four possible answers were correct.

To become familiar with the task and the method of response, each subject ran trials consisting of 27 of the easiest stimuli (the 320 dot 0.5*s*-height stimuli). Subjects ran trials until accuracy was at least 85% correct (approximately 100–130 trials).

Results

Accuracy data. All subjects reported that they perceived a 3D surface the first and every subsequent time they viewed the high numerosity displays. With low numerosities, the dots were perceived in approximately their correct positions in 3D space, but there were too few dots to give the illusion of a continuous surface or to discriminate unambiguously between alternative responses. The very limited practice served merely

to teach the subjects to name the perceived shapes without having to refer to drawings.

The results of Experiment 1 are summarized in Figure 2. Each response was scored as correct only if both the shape and the rotation direction were correct and consistent. Thus, if *u*+−0*l* was the display, responses *u*+−0*l* and *u*−+0*r* were correct. Every other response was incorrect. There were occasional responses with the correct shape and the incorrect rotation direction (66 such errors, 4.5% of all responses, 10% of all errors). Subjects later indicated that most of these were a result of forgetting the direction of rotation before the response was completed, rather than from a truly misrotating percept. Nevertheless, such responses were treated as incorrect.

As expected, accuracy improved both with the numerosity and with the amount of depth displayed. There were signs of a ceiling in performance as numerosity increased. For two

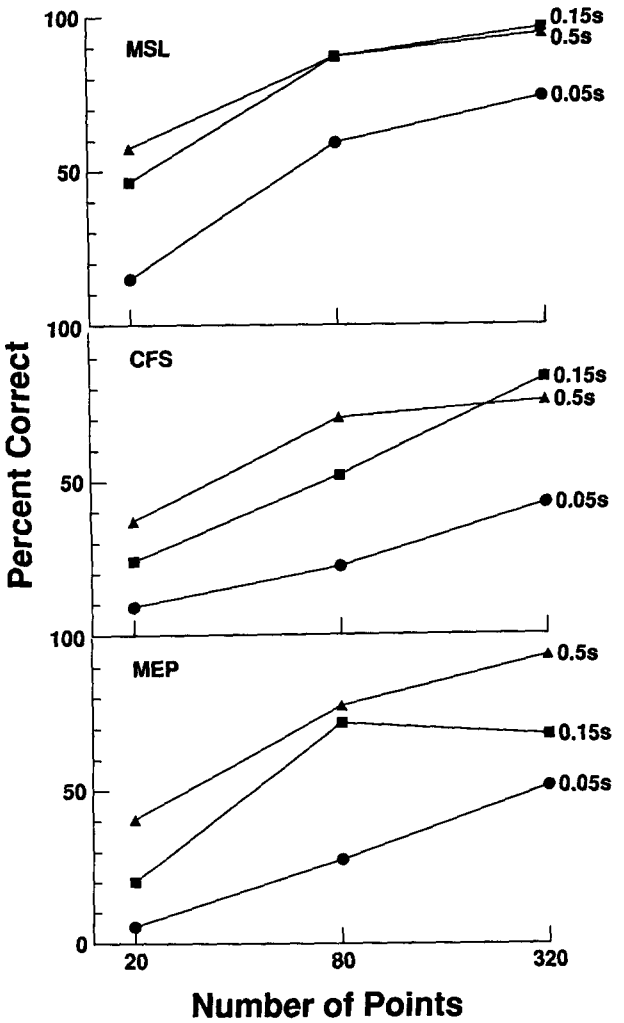

Figure 2. Performance on the shape identification task as number of points in the simulated shape was varied. (The parameter is the height of the bumps relative to the length of a side. Each panel represents data from a different subject. Performance increased with both numerosity and bump height.)

subjects, for 320 point displays, the curves crossed, and the middle-range depth extent (0.15*s*) was as good or better than the large 0.5*s*-depth extent. An analysis of variance was computed treating numerosity, height, and subjects as treatments, and shapes/rotations as the experimental units. Both numerosity and degree of depth were highly significant ($p < .0001$), with $F(2, 106) = 119.0$ and $F(2, 106) = 102.9$, respectively. Subjects differed significantly from one another, $F(2, 106) = 33.5$, $p < .0001$. The three-way interaction was significant, $F(8, 424) = 2.6$, $p < .01$, indicating that the interaction of height and number differed among subjects (see Figure 2). No two-way interactions were significant.

Error analyses. A confusion matrix was computed, pooled across subjects, the nine conditions, two rotation directions, and two possible designations of each shape or depth reversals (it was thus a 27 × 27 = 729 cell matrix). Table 1 is a summary of these identification errors. Descriptions are given for seven common error types, one uncommon error type and a miscellaneous category. If a bump and a depression were present in the display, and only one of the two was indicated by the subject, this was called a *missed feature error.* If the bump and depression are of equal extent on the base plane (e.g., *u*+−0), then this was called a *missed equal size feature.* If they were of unequal extent, and the smaller of the two was not reported, this was categorized as a *missed smaller feature.* Any display that contained only one depth sign (such as *u*+00) and was reported as containing both depth signs (e.g., *u*0+−) was categorized as *report two depth signs when there was only one.* For any given row in the table, the second column presents examples of errors of that row type. The third column lists the number of cells in the confusion matrix that correspond to an error of a given type, and the fourth column provides the total number of errors that occurred over all cells of that type. The last column is the average number of errors per cell in cells of that type, computed as the ratio of the number of trials indicated in Column 4 divided by the number of cells in Column 3. In total, there were 586 errors; divided by 702 error cells this yields 0.83 errors per cell on the average. A ratio greater than 0.83 in Column 5 of Table 1 indicates an error type more common than the average, a smaller number indicates a less common than average error type.

The bottom row of the table provides summary information. The first seven error types listed had ratios well over this value and thus were more common than other errors. The *report two depth signs* ... error type is an example of an exceedingly uncommon error.

The quantity of data collected was not sufficient to enable us to confidently draw many specific conclusions from the error data. The hypothesis that errors are distributed uniformly across the nine error classes was easily rejected, $\chi^2(8, N = 586) = 1{,}032$, $p < .001$. It appears that four types of errors were the most prevalent. Large single bumps were highly confusable, especially the subtle difference in shape that distinguishes *d*+++ from *u*+++, but also that distinguishes between *d*+++ and *d*0++, and so on. Errors were made in horizontal location of the shape within the ground (e.g., *d*0+0 was reported as being *u*+00, or *d*++0 as *u*+0+). Errors were also made in judging the width of the bumps

Table 1
Summary of Identification Errors, Pooled Over Subjects, Bump Heights, Dot Densities, Rotation Directions, and Depth Reversals

Description	Examples	Number of cells	Number of errors	Ratio[a]
Small distortions of large bumps	*u*+++ interchanged with *d*+++	2	29	14.5
Incorrect bump width, correct location	*u*0++ interchanged with *d*+00	4	34	8.5
Missed smaller features	*u*++− reported as *u*++0	6	30	5.0
Diagonal bump reported as large bump	*u*++0 reported as *u*+++ or *d*+++	8	23	2.9
Missed equal size feature	*u*+0− reported as *u*+00	12	29	2.4
Incorrect diagonal bump size	*u*++− reported as *u*+0−	8	16	2.0
Small horizontal location error	*u*+00 interchanged with *d*0+0	16	27	1.7
Report two depth signs when there was only one	*u*+00 reported as *u*+−0	168	40	0.24
Other errors	—	478	358	0.75
All errors	—	702	586	0.83

[a] Total number of indicated error responses divided by total number of applicable cells (Column 4/Column 3). A ratio greater than 0.83 indicates a type of error that is more common than average.

(e.g., *d*+00 was reported as *u*0++). Finally, for displays for which both a bump and a concavity were present, occasionally one of the two was not noticed. It is interesting to note that in every case of this type of error (the missed smaller features and missed equal-size features of Table 1, and the less common missed larger features), the response was of a single bump toward the observer. In other words, in the presence of a perceived convexity, a concavity is occasionally missed, but not the other way around. On the other hand, when only one nonzero depth was present (a single bump or concavity), it was very rare for subjects to give a response containing multiple depth signs.

When the confusion matrix was broken down by experimental condition, the amount of data was rather low. Nevertheless, a few interesting trends were evident. First, all seven common error types (the first seven rows of Table 1), remained common in all experimental conditions. As the task became more difficult, the types of errors subjects made remained "sensible." Second, the first two error types, although common in difficult conditions (low height or low numerosity), became even more common in easier conditions. As the shape impression improved, the subjects were able to eliminate other possible shapes and then were more likely to err by choosing the most similar incorrect shape. The distinction between *d*+++ and *u*+++ was very difficult to make even when the perception of depth was quite compelling and well sampled. The *report two depth signs* . . . error type was uncommon in all conditions, but there appeared to be a trend for this error type to become more common as numerosity increased.

Experiment 2: Texture Density

Several cues may lead to correct shape identification in the KDE task. One cue is dynamic changes in texture density. The shapes are generated in such a manner that, head-on (i.e., viewed with the object base plane in the picture plane), the expected local dot density across the display is uniform. By itself, the initial frame has no shape information whatsoever. As the shape rotates, areas in the display become more dense or sparse as the areas in the shape that they portray become more or less slanted from the observer. Theoretically, the observer could use this cue from subsequent frames after the first to determine the shape. Because we are interested in structure from motion, the changing texture density adds a cue in addition to the relative motion cue. In Experiment 2, we compared three conditions: (a) Both the motion and density cues were present as before; (b) only the motion cue was present—dot lifetimes were varied in such a way as to eliminate the density cue by keeping local average dot density constant across the display; and (c) only the density cue was present—the relative motion cue was eliminated by reducing dot lifetimes to just one frame.

Method

Subjects. Three subjects were used in the study. One was an author of this article; two were graduate students naive to the purposes of the experiment. Two had corrected-to-normal vision; one subject (CFS) had vision correctable only to 20:40.

Displays. The displays were generated in a manner similar to Experiment 1. The same lexicon of 53 shapes was used. The flat ground surrounding each shape was extended horizontally by 20% and was later windowed to the same 182 × 182 pixel, 4° square, so that the sides of the displays no longer oscillated with the rotation. Instead, points appeared and disappeared at the edges of the window. For each shape, an instantiation of the shape was made with 10,000 points and with the large 0.5s-bump height of Experiment 1. Displays for each of the three experimental conditions were made by randomly subsampling points from this rotating 10,000-dot shape.

Control condition: Motion and texture cues. The control condition had both the relative motion and changing texture density cues. A small random subsample of points was chosen, so that approximately 320 points were visible through the 4° square window. The subsample of points was rotated and projected as before, and then clipped so that only those points within the window were displayed. This condition was identical to the easiest condition of Experiment 1 (0.5*s*, 320 dots) except for the windowing (and the lower dot contrast described later). Examples of the density cue available in these displays are shown in Figure 3.

Only motion cue. This main experimental condition removed the changing texture density cue (Figure 3). The 4° × 4° square window was treated as consisting of a 10 × 10 grid of subsquares. Texture density was kept uniform by forcing each subsquare to contain exactly 3 points in every display frame. Thus, there were exactly 300 points visible in every frame. On the first frame, 300 of the 10,000 points were randomly chosen, subject to the constraint that exactly 3 points were chosen in each subsquare. On each subsequent frame, the 10,000 points were rotated by the proper amount. Then, for each of the 100 subsquares, the points (of the 300) that then appeared in each subsquare were counted. If more than three occurred, points were randomly chosen and marked as no longer displayed, until the number of displayed points in that subsquare fell to 3. If less than 3 points in a grid square were displayed, then more points were randomly chosen (from the 10,000) that would then appear in that subsquare to bring the total back up to 3. In this condition, dot density remained uniform throughout the display. Points were deleted or reinstated only as needed to keep the density uniform. Although variations in texture density were noticeable in the control displays, the exclusion of the density cue did not seriously disrupt the correspondence of the majority of the points: Most points remained displayed for 10 frames or more during the 30 frame display.

The amount of scintillation was small. The average change (one half of total dot additions plus deletions) between two frames was 16; for 300 dot displays this was 5.3% scintillation. (The highest between-frame scintillation was 8.3%.)

Only texture density cue. The relative motion cue was removed in this condition leaving the changing texture-density cue intact. For each frame in the display, 320 of the 10,000 points were randomly chosen. This happened independently on every single frame, subject to the constraint that no point ever appeared in two successive frames. Thus, no relative motion cues were available in these displays, which looked like dynamic sparse random dot noise. On the other hand, because the points were chosen randomly from the 10,000 points, they had the same expected texture density as the 10,000 points on each frame, and indeed became more dense and sparse in exactly the same fashion as in the first experimental condition (as illustrated in Figure 3).

There were 53 possible shapes and three experimental conditions, resulting in 159 display types. Two different displays were made of each display type of the flat shape, and one display was made for all other display types. There were thus 162 displays. They were displayed as bright green dots on a green background of lesser luminance. The display background luminance was 31 cd/m^2. Each dot added an additional 13 μcd, viewed from a distance of 1.6 m. All other display characteristics were the same as in Experiment 1.

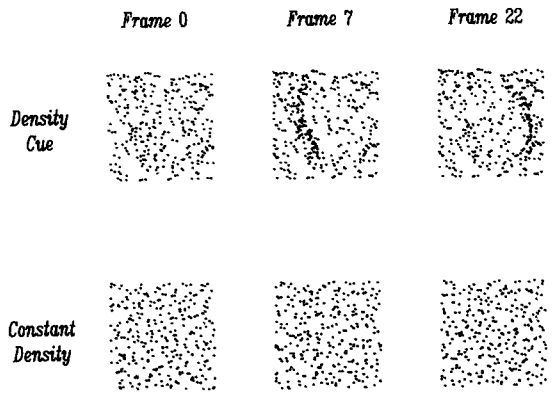

Figure 3. The dynamic density cue. (Three frames are shown from a display corresponding to u+0+r, a bump extending from the top center to the lower right. The upper row shows frames with the density cue. The lower row illustrates the effectiveness of removing the density cue in the motion-only condition.)

Apparatus. The apparatus was the same as in Experiment 1. Only the green channel of the Conrac display monitor was used.

Viewing conditions. The viewing conditions were identical to Experiment 1.

Procedure. There were 11 experimental conditions: the 3 described previously (motion and texture, motion only, texture only) and 8 others that will be reported elsewhere. There were thus a total of 594 displays, including the 162 displays of the 3 conditions reported here. These were presented in a mixed-list design in four sessions of 1 hr each. Otherwise, the procedure was identical to Experiment 1.

Results

Density cue. The results are shown in Figure 4. For two subjects (MSL and CFS), elimination of the changing density cue did not alter performance. For the third subject (JBL), performance dropped from 81.5% to 68.5% after the density cue was eliminated. However, it was not clear whether this small performance change was due to the elimination of the density cue itself or the introduction of scintillation (dot noncorrespondences) by the process of eliminating density cues. For two subjects (CFS and JBL), the elimination of the relative motion cue in the density only condition dropped performance to levels that did not differ significantly from chance. For the third subject (MSL), performance with the density cue alone was significantly above chance, although well below performance for conditions in which the relative motion cue was available.

In the condition in which only the changing dot density cue was available, the displays did not look 3D. The only subject (one of the authors) who was able to perform significantly above chance in this condition was highly familiar with the construction of the displays. For any given shape and rotation direction, clumps of higher density appeared first on one side of the display, and then later on the other side, as the object was rotated an equal amount in both directions from the initial face-forward orientation through the course of the 30-frame display. Performance was a matter of noting the positions in the display at which high density occurred,

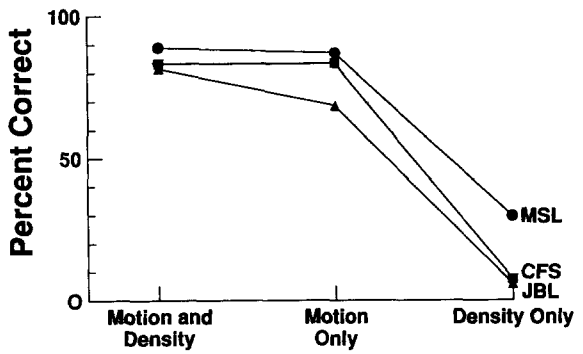

Figure 4. Percentage of correct shape-and-rotation identifications for the three cue conditions of Experiment 2. (Data are shown for 3 subjects.)

on which side of the display they occurred first, and the 2D shape of the texture clump. Then, a response was chosen that was most consistent with this information. This was a highly cognitive task, and it took far longer to respond in this condition as a result.

Changing dot density was neither a necessary nor a sufficient cue for the perception of 3D shape with these displays. However, when the density cue was available with motion cues, the density cue may have been used by one of three subjects to slightly improve his responses. When the density cue was the only cue, another one of three subjects was able to improve his response accuracy to significantly above chance. These results point out the importance of removing artifactual cues from kinetic depth displays.

Scintillation cue. In the constant density condition, one might argue that the subject was indirectly provided with shape information by the amount of scintillation (dot noncorrespondence) in different areas of the display. Local scintillation could potentially be used by a subject (just as density information was useful to one of three subjects in the density-only condition).

The relation between local scintillation in these displays and local density (and thereby, ultimately, local shape) in the control displays is not simple. Points are deleted or added only when necessary to keep the number of points in a given locale constant. The number of points that will be added (or deleted) is thus proportional to the local rate of change of texture density. The difficulty in computing shape from scintillation is that subjects are poor at judging the degree of scintillation in a pattern, other than differentiating some scintillation from no scintillation (Lappin et al., 1980). And it is even more difficult to determine whether scintillation is due to points being added or to points being subtracted, that is, to determine the sign of the change of texture density.

We further investigated the possibility that scintillation might have been a useful cue, in an informal experiment. Various amounts of irrelevant scintillation (in the form of fresh, randomly occurring dots in each frame) was added to all areas of each frame. With added scintillation that was 10 times more than that produced by the density removal program, the quality of the image was greatly impaired. But the ability to discriminate shapes seemed to be unimpaired. This means that scintillation is relatively unimportant: Large amounts do not greatly impair the display; small amounts are not necessary to perceive KDE because, when they are masked by large amounts of scintillation, performance hardly suffers.

In displays similar to those of Experiment 2, restricting dots to have lifetimes of only 3 frames was another operation that generated large amounts of scintillation. KDE identification performance remained high even though the amount of scintillation was large and varied randomly throughout the display and from frame to frame (Dosher, Landy, & Sperling, in press; Landy, Sperling, Dosher, & Perkins, 1987). All in all, the difficulty subjects had in estimating the amount of scintillation in the first place and the subsequent difficulty of any computation for estimating shape from scintillation made it unlikely that scintillation played a significant role. We conclude that density-related shape cues are eliminated in the motion-only displays.

Experiment 3: Equivalent 2D Task

Because of the large set of shapes, the systematic way in which it was constructed, and the large set of possible responses, it appears difficult to perform accurately in this task without a global perception of shape. Indeed, except in the case of the density-only displays of Experiment 2, all of our subjects reported perceiving a global shape and basing their response on this global shape percept. Nevertheless, one of our most serious objections to previous studies of KDE was that the subjects could have performed the experimental tasks without a global perception of shape by using minimal, incidental cues. Because our set of shapes was finite (53 shapes), there were indeed potential artifactual strategies; however, because each realization of a shape was composed of different random dots, we were unable to discover any simple, minimal computation for our task. The simplest computation was equivalent to what we believe the KDE computation itself to be.

To study alternative mental computations that might yield correct responses in our KDE task, we developed a new display that did not produce the 3D depth percept of KDE but that was as equivalent as possible to the KDE display in other respects. To perform correctly with the new display, the subject would have to perform a computation that was equivalent to the KDE computation except in that it is performed by some other perceptual/cognitive process, a process that did not yield perceptual depth. We call such a computation a *KDE-alternative computation.*

Suppose that a subject chose to perform the shape identification task by measuring instantaneous velocities at only a small number of spatial positions and making this velocity determination at only a single moment during the motion sequence, for example, a moment at which velocities were the greatest. A high velocity indicates a point far forward or far behind the base plane. Opposite velocities indicate points at opposite depths. Using these simple principles, it is obvious that velocity measurements at six positions, the corners of both triangles used in specifying the shapes, would be sufficient to identify the shapes. Fewer measurements of velocity made at intermediate points would suffice for identification of our restricted set of stimuli, but they would involve unrealistically complicated computations that were specific to this stimulus set.

In Experiment 3, we evaluated a computation for shape reconstruction based on a strategy of making six simultaneous local velocity measurements at the points that corresponded to the possible depth extrema in our stimulus set.

Method

Choosing motion trajectories for display. In the shape identification task (Figure 1), suppose one were to track a single point on the surface of the shape throughout the course of the display. Initially the point is at position (x, y, z), where x and y are the horizontal and vertical image plane axes, respectively, and z is the depth axis. As in Experiments 1 and 2, assume that the shape is rotated about the y axis according to $\theta(m) = \pm 25 \sin(2\pi m/30)$, where m is the frame

number. Under parallel projection, the motion path of the point is purely horizontal:

$$x(m) = r \cos\left\{\frac{2\pi}{360}\left[\theta_0 \pm 25 \sin\left(\frac{2\pi m}{30}\right)\right]\right\},$$

where $r = (x^2 + z^2)^{1/2}$, and $\theta_0 = \tan^{-1}(z/x)$ degrees.

If the subjects were to apply the local motion strategy to the shape identification task, they would need to measure and categorize local velocity for six such motion paths simultaneously. In Experiment 3, the subjects were presented directly with stimuli containing six moving patches and they were requested to categorize the local directions of motion.

Displays. Each display was based on a particular shape from the shape identification task. Each of the six motion paths portrayed in the display was based on a motion path followed by a critical point on the surface of the shape, as just described. The six critical points were the projections onto the surface of the six points originally used to generate the shapes (see Figure 1A, *u* and *d*). The motion paths were based on the shapes with the largest heights ($h = 0.5s$, where s is the width of the visible background plane).

The displays were intended to force subjects to use the strategy of simultaneously measuring six velocities, without any possibility of recourse to using perceived 3D shape. Each display consisted of six patches of moving random dots (Figure 5). The dots within a patch all moved with the same velocity, and patches were spatially separated, so that there was no perception of depth. The outline squares of Figure 5 were not directly visible to the subject. They acted as windows through which planes of moving random dots were seen. Due to a setup error, dot density in Experiment 3 was slightly less (0.83 of rather than equal to) than the density used in the constant density condition of Experiment 2. (This density difference was so small that it went unnoticed at the time.)

Response mapping. There were two rows of three patches of moving dots. Figure 5 indicates the correspondence of patch position to where that patch's motion is visible in the original shape displays. Spatial positions in Experiments 2 and 3 were essentially similar

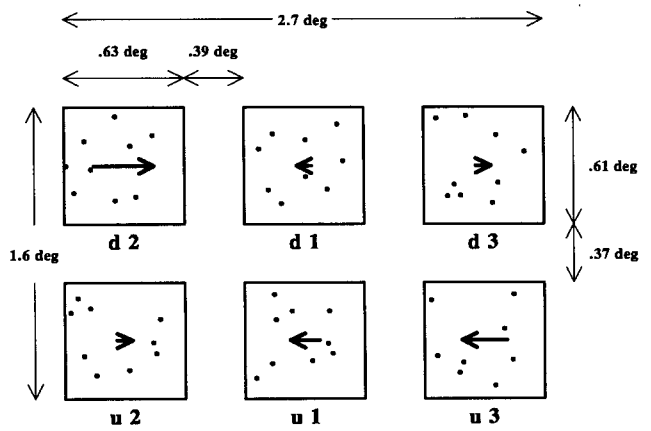

Figure 5. Spatial layout of the stimuli used in Experiment 3. (The squares represent windows through which fields of moving random dots were seen. The outline of the windows was not visible to the subject. The label under each window denotes the position in the shape, as in Figure 1A, that controlled the motion portrayed in that window. For example, the motion path of all the random dots seen in the upper middle window was the same as that taken by the point in a shape display of Experiment 1 that was initially above position *l* in the *d* triangle shown in Figure 1A.)

except that the middle positions in each row of Experiment 2 displays were interchanged to create the Experiment 3 displays. This was done in order to make the response easier for the subjects. With the KDE shape displays, the subject decided whether the three important points were those of the *u* or *d* triangle, and then categorized the height at each of the three corners of that triangle. In the corresponding motion task, the subject decided whether the top or bottom row of patches was most important, and then categorized the motion path of each patch in that row.

For points at a reasonable height above the base plane, the 2D motion path was quasisinusoidal. That is, points moved to the left, then to the right, then returned leftward to their starting position (or right, then left, then right). Points with a larger initial z value moved faster. The extreme z values generated the highest speeds, and these always lay above the vertices of the base triangle used to generate the shape. This meant that subjects could solve the motion task by first judging which row contained the fastest speed, and then, for that row, categorizing the motion in each of the three patches about halfway through the course of the display time. Each patch was to be labeled as moving quickly to the left (*l*), quickly to the right (*r*), or slowly, if at all (0). Note that points in the other row also moved in a quasisinusoidal manner, but more slowly than the maximum speed in the relevant row.

One possible response was, for example, *ulr*0. This response would indicate that the fastest speeds were in the upper row: the upper-left patch moved right, then left, then right, the upper-middle patch moved left, then right, then left, and the upper-right patch was moving slowly. There were 54 possible responses (2 rows, 3 possible motion categories for each of the three patches in that row). Because *u*000 and *d*000 denoted the same display (one in which all patches were moving slowly), this yielded 53 distinct display types, corresponding to the 53 distinct shape-and-rotation display types in the shape-identification experiment.

There were 53 possible shapes. With 2 exemplars of the flat shape, and 1 for all other shapes, this yielded 54 displays. Motion displays were displayed as bright white dots on a gray background. The display background luminance was 15.6 cd/m^2. Each dot added an additional 24.3 μcd, viewed from a distance of 1.6 m. All other display characteristics were the same as in Experiment 1.

Apparatus. The apparatus was the same as in Experiment 1, except that a monochrome U.S. Pixel PX15H315LHS monitor with a fast, white phosphor was used.

Viewing conditions. Stimuli were viewed monocularly with goggles; a circular aperture restricted the field of view. Luminance outside the aperture was approximately equal to the background luminance on the CRT, which was 15.6 cd/m^2. Stimuli were viewed from a distance of 1.6 m. After each stimulus presentation, the subject keyed responses using response buttons, and visual feedback was given on the CRT. The room was dark, but light adaptation level was controlled by the CRT background and the illumination of the occluding screen.

Procedure. A block of trials consisted of 108 trials. Each of the 54 displays was viewed twice in random order. For the stimuli based on the flat shape, two possible answers were correct (*u*000 and *d*000). For all other stimuli only one answer was correct.

Subjects were told precisely the correct strategy to use. They were told that they would see six patches of moving dots. They were to determine which row contained the patches with the fastest motion (either the upper row, designated *u*, or the lower row, designated *d*). For that row, subjects were to categorize the motion in each of the three patches in that row as measured about halfway through the course of the display time. Each patch was to be labeled as moving quickly to the left (*l*), quickly to the right (*r*), or slowly if at all (0). After each response, the correct answers were displayed as feedback. Other details of the procedure were identical to Experiment 1.

Subjects. Two subjects were used in the study. One was an author of this article, one was a graduate student naive to the purposes of the experiment. Both had corrected-to-normal vision. Subject MSL ran a single block of 108 trials. Subject JBL ran three blocks of 108 trials.

Results

Subject MSL scored 90.7% correct on a single block of 108 trials. In the three blocks of trials run by subject JBL, the scores were 58.3%, 75.9%, and 88.0%, respectively. Indeed, after a little practice, performance was quite good, equal or slightly better than performance in the easiest conditions of Experiments 1 and 2, which had a comparable dot density and range of velocities.

There were too few trials to make an in-depth analysis of error data. However, the most frequent motion response errors corresponded to the two most frequent KDE errors in Table 1 (small distortions or mislocalizations of large bumps). For example, 8 out of the 10 errors made by MSL were analogues of these two error types. Examples: *ulll*, a triple "up" bump was reported as *dlll*, a triple "down" bump; *ul0l* was reported as *d0l0*; a double bump was mistaken for a single bump in the same location (see Figure 1). Indeed, these results are not surprising because the velocities involved in Experiments 1 and 3 were similar. It seems likely that a very large number of trials would be required to find any significant differences in the error patterns in Experiment 3 and those in Experiment 1.

Discussion

We have introduced a new objective task for measuring the perceptual effectiveness of the kinetic depth effect: shape identification. With the current lexicon of shapes, it measures whether the subject can globally determine precisely which areas are in front of the ground and which areas are behind the ground. We consider here some possible objections to and some issues raised by our results.

Cues to Structure From Motion: Optic Flow or Interpoint Distances?

In the displays of Experiment 2, in which dot density was controlled, subjects solved the shape identification task even though no single frame contained any information that could have been used to infer shape. For these stimuli, at least two frames were needed to infer shape. By definition then, the only possible cues were motion cues.

There are at least two possible motion cues to depth: optic flow and changing interpoint distances in the displays. That is, subjects could be deriving shape from a global optic flow field (instantaneous velocity vector measurements across the field) or from measurement of interpoint distances of particular dots over two or more frames. Models of the KDE have been based on both optic flow (Koenderink & van Doorn, 1986) and on interpoint distances (Hildreth & Grzywacz, 1986; Landy, 1987; Ullman, 1984). To a certain extent, it is possible to differentiate between these models by creating displays in which dots have lifetimes of only two frames. In such displays, a global optic flow field is available (although noisy), and 3D structure could, in principle, be computed from the flow field. Alternatively, some subset of the points could have been used to compute a 3D object based on interpoint distances. However, the particular object changes rapidly because within two frames all points have been replaced by entirely new points, uncorrelated with those of the preceding frames. It turns out that subjects are quite adept at the shape identification task with such displays (Dosher et al., in press; Landy et al., 1987). This, and related results, are taken as strong evidence against the interpoint distance models (Dosher et al., in press; Landy et al., 1987). Together with the results of the present experiment, in which changing density is eliminated as an alternative, this leaves motion flow fields as the necessary and sufficient cue for KDE in moving-dot displays. Whether interpoint distances or other motion cues are ever perceptually salient remain open questions.

Multiple Facets of the KDE

We have previously argued (Dosher et al., 1989; Landy et al., 1985) that measurement of the full effect of stimulus manipulations on the KDE requires several subject responses in order to describe fully the richness of the percept. These responses included judgments of coherence (whether the multidot stimulus coheres as a single object), rigidity (does the object stretch?), and depth extent (what is the amount of depth perceived?). These different aspects of the percept are partially correlated, but they can be decoupled by suitable display manipulations. For example, with some subjects, the addition of exaggerated polar perspective to a display increases the perceived depth extent even as it decreases perceived rigidity.

In the current experiments, this richness of the KDE percept was not explored. We measured the extent to which the display was effective in creating a global sensation of depth, and hence supported objective shape identification. Other aspects such as depth extent or rigidity were not measured. The difference between the three depth conditions was immediately obvious to subjects, and increasing the depth extent displayed (within certain limits) did improve performance, but we did not measure perceived depth extent.

Although perceived rigidity was not explicitly measured, nonrigid percepts were spontaneously reported by subjects. One particular example was very common. Shapes with both bumps and concavities (e.g., *u++−−*) were occasionally seen in a nonrigid mode. Rather than seeing one area forward, another one back, and the whole thing rigidly rotating, observers perceived both areas as being in front of the object ground and rotating in opposite directions (this percept looks rather like a mitten with the thumb and finger portions alternately grasping and opening). This particular nonrigid percept occurred most often when the number of dots was large and the depth extent was at its largest. In this stimulus condition, with mixed-sign shapes, it is clearly visible that the two bumps cross (in the rigid mode, one sees through the bump to the concavity behind it when they cross). This is an example of a failure of the "rigidity hypothesis" (Adelson,

1985; Braunstein & Andersen, 1984; Dosher, Sperling, & Wurst, 1986; Schwartz & Sperling, 1983; Ullman, 1979), because a stimulus that has a perfect rigid interpretation is perceived as nonrigid. (It should be noted that the nonrigid interpretation also is a veridical 3D interpretation that is consistent with the 2D stimulus; it happens not to match the required response mapping.) These stimuli are multistable, yielding more than two possible stable percepts. In our experiments, when subjects perceived a nonrigid object, they were required to compute the name of one of the possible rigid objects that was consistent with what they perceived.

Relations to Previous Empirical Studies

We found that shape identification performance increases with the number of dots displayed and the extent of depth portrayed. Neither of these results is surprising. The numerosity result is an extension of previous, more subjective, measures of the depth perceived in simple KDE displays (Braunstein, 1962; Green, 1961). Increasing the number of dots provides the observer with more samples of the motion of the shape portrayed. Increasing depth extent increases the range of velocities used. Both manipulations increase the observer's signal-to-noise ratio in the task, in which noise sources may be both external (such as position quantization in the display and sparse shape sampling) and internal.

What is Computed in KDE?

Within measurement error, subjects performed equally well in the motion judgment task of Experiment 3 and comparable KDE tasks of Experiments 1 and 2. Further, the most common confusion error was the same in all experiments. And there is every reason to suppose that, if more data were available, the less common errors also would be highly correlated. In brief, we have succeeded in creating two equivalent tasks for classifying stimuli into 53 shape categories: One is solved by a KDE mechanism that yields a perceived 3D shape, and the other is solved by a motion perception mechanism that yields a perceived pattern of 2D motions. What does this imply about the mechanism of KDE and about the technology of KDE experimentation?

Although the specific nature of the perceptual algorithm that extracts 3D structure from 2D motion has not yet been established, it is reasonable to expect that it ultimately will be. Whatever the computation, the equivalent computation could, in principle, be carried out by some other system that was supplied with the same raw information, in this instance, the optical flow fields. In Experiment 3, we demonstrated that the measurements of the optic flow fields at six points provide sufficient information for the shape categorization task. When the optic flow at these locations is provided to observers in a response-compatible format, they can use this optic flow information to categorize the stimuli in perceived 2D just as efficiently as when they categorize KDE stimuli in perceived 3D. What is special about extracting structure from motion is not the informational capacity of the KDE system, but the perceptual capacity for extracting the relevant information and providing it perceptually as 3D depth.

For extracting structure from motion, the relevant information is optic flow. This was demonstrated in Experiment 2 (in which the residual nonflow cues were eliminated) and by experiments in which dots were given maximum lifetimes of only two (or three) frames so that correspondence cues were weakened and only optic flow cues survived (Dosher et al., in press; Landy et al., 1987). The relevant information in our particular shape discrimination task is the set of local velocity minima and maxima in the optic flow and their approximate shape. A reasonable assumption about the structure-from-motion computation is that the perceptual system automatically locates these maxima and minima, extracts the velocities, and transforms them into perceived depths. (Relative velocity has long been recognized as an extremely potent depth cue [e.g., Helmholtz, 1910/1924, p. 295ff; Rogers & Graham, 1979] and undoubtedly is a critical component of KDE.) When the relevant areas of optical flow are extracted instead by our display processor and presented to the subject as isolated patches, the subject is still able to classify the velocity in the patches, but the automatic perceptual conversion of velocity into perceived depth is inhibited. Nevertheless, the extracted velocity information is sufficient to enable accurate classification of the stimuli when a response-compatible format is made available.

Figure 6 illustrates the processes that are assumed to be involved in object recognition via the KDE. From the stimulus, the subject extracts a 2D velocity flow field. The KDE is the process whereby 3D depth values are extracted from the flow field. These depth values are combined with other shape and contour information from the stimulus to yield a 3D object percept which then forms the basis for the subject's response. A KDE-alternative computation is one that uses the same stimulus and velocity flow field, but circumvents the KDE computation by deriving the required response directly from the flow field. Experiment 3 demonstrated that a KDE-alternative computation would be possible in principle if the subject could extract the velocities at the six most relevant locations.

In transforming flow-field velocity into perceived depth, there is an inherent ambiguity in sign: A given velocity can equally well indicate depth toward or away from the observer. This ambiguity is inherent in the optics of the display and reflected in our scoring procedure. However, the perceptual system tends to resolve the ambiguity consistently in nearby locations. On those occasions in which it does not (e.g., when it interprets leftward motion as closer in one display area and as further in another), the display appears to be grossly nonrigid. The likelihood of consistent depth interpretation has been studied by Gillam (1972, 1976) and probably can be modeled by locally connected cooperative-competition networks (see Sperling, 1981, for an overview of cooperation–competition in binocular vision and Williams & Phillips, 1987, for an example of cooperation in motion perception).

KDE-Alternative Computations

It is useful to distinguish three kinds of computations: KDE, KDE-alternatives, and artifactual non-KDE computations. The KDE computation is an automatic perceptual computa-

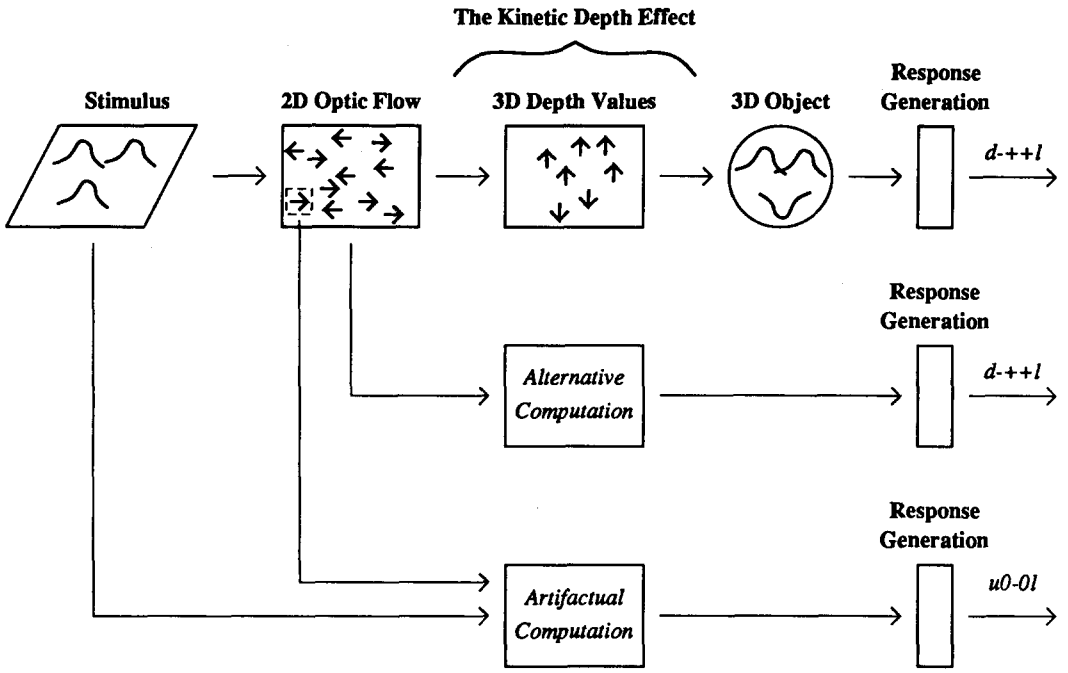

Figure 6. Flowchart for KDE, KDE-alternative, and artifactual computations. (From the stimulus, the following are assumed to be computed in sequence: 2D velocity flow field, 3D depth values [KDE computation], a 3D object representation [which in this instance happens not to correspond perfectly with the object represented by the stimulus], and the required response sequence. The KDE-alternative computation computes the required response sequence directly from the 2D optic flow without an intermediate stage of perceived 3D depth; that is, it simulates the KDE computation in another part of the brain. An artifactual computation uses incidental stimulus cues or motion cues from only a small part of the stimulus to arrive at a response.)

tion made, in the case of our stimuli, on velocity flow fields, and it results in perceived depth (a 3D percept) at those visual field locations where it is successful. A KDE-alternative computation is a computation on velocity flow fields similar to the KDE computation except that it is made consciously in some other part of the brain. It results in a knowledge of the correct response, but it does not yield perceived depth: The field is perceived as flat. An artifactual, non-KDE computation uses an incidental property of the display to compute the correct response, and the computation may be quite unrelated to the KDE computation. For example, the various objective studies of KDE that we considered in the beginning of this article all were vulnerable to computations that used only a small portion—in some instances only the movement of a single dot—of the stimulus information that would have been required by a KDE computation.

Of the five studies reviewed in the beginning of this article, the possible artifactual computations involved 1 dot (one study), 2 dots (two studies), and other cues (two studies). The problem is purely technical; the possible artifactual computations are quite different from KDE computations. There is a great risk of admitting an artifactual computation when the set of possible stimuli is small and when the required KDE computation itself is relatively simple. Even though subjects in these studies may have perceived KDE depth, a simple 2D strategy would have improved response accuracy. Although some of these procedures could have been improved, we deemed it better, from the outset, to use a large set of stimuli that can be identified only after a relatively elaborate KDE computation. What distinguishes the present task from prior tasks is that they admitted artifactual computations that were shortcuts to the correct response; the present alternative computation is an equivalent computation to KDE.

With respect to KDE-equivalent computations, we can ask two questions: Do they ever occur, and if they do, how can we be sure that they do not always occur? To demonstrate that a KDE-equivalent computation can occur we first have to know what the KDE computation itself is, and then to perturb the stimulus so that the automatic KDE computation cannot occur. In our experiment (and probably more generally), the essential KDE computation is the discovery of local velocity minima and maxima, and the consistent depth labeling of these minima and maxima. In Experiment 3, six stimulus areas around the velocity extrema were extracted from the KDE stimulus, and (in order to avoid the automatic KDE computation) they were presented as isolated squares. The subjects were able to label these areas consistently with respect to velocity (not depth, because the display was perceived as flat). Thus, subjects performed a KDE-equivalent task by means of a KDE-equivalent computation. Furthermore, the pattern of errors in the equivalent task corresponded to the previous error pattern in the KDE task. Although there are necessarily some differences between the KDE stimuli and the alternative stimuli, our strong result makes it clear

that, along with artifactual computations, the possibility of a KDE-alternative computation has to be considered in interpreting KDE experiments.

Artifactual computations are most easily discriminated from KDE computations by varying stimulus parameters. Stimulus cues that might support an artifactual computation are removed, masked or are rendered useless by irrelevant variation. If response accuracy survives, we have increased confidence that it is based on a KDE computation.

KDE and KDE-alternative computations use the same stimulus attributes; they differ in where in the brain the computation is made. Two tools for discriminating between these computations are *introspection* and *dual tasks*. For example, all subjects, without conscious effort, immediately perceive our KDE stimuli as solid 3D objects. When subjects honestly report that they perceived 3D depth in dynamic KDE stimuli, by definition, they have performed a KDE computation. The problem is that KDE may not be the only computation being performed. For complex stimuli such as ours, however, it is hard to imagine that a subject could be performing a useful alternative computation without awareness. Indeed, the discovery of an alternative computation for KDE is the structure-from-motion problem, and the solution proposed in Experiment 3 may be the first workable solution for stimuli of this type. It would be remarkable if subjects, even sophisticated subjects, discovered the solution in the course of viewing the stimuli. Still, even in this case, but especially with simpler stimuli, it would be better to use a formal procedure to exclude alternative computations. This requires, for example, (a) isolating the alternative computation, as in Experiment 3, (b) finding a concurrent task or similar manipulation that selectively interferes with the alternative computation relative to the direct KDE-computation, and (c) using the modified or dual tasks with the original stimuli.

An alternative KDE computation is analogous to an alternative stereoptic depth computation that is carried out by monocularly examining the left and right members of a stereogram. When stimuli are designed to take advantage of the exquisite sensitivity of stereopsis, an alternative monocular computation that uses remembered disparities is not feasible, even though it may be learnable in special cases. The same is undoubtedly true for KDE and alternative KDE computations: For complex KDE stimuli, viewed briefly, the alternative computation is simply out of the question. However, the problem of interpreting experimental results has not been alternative KDE computations but artifactual non-KDE computations. The best way to avoid subsequent problems of interpretation is to use complex stimuli, like the 53-shape stimulus set used here, that are matched to and challenge the ability of the human KDE computation.

Summary and Conclusion

A new shape identification task for measuring KDE performance is proposed. With its lexicon of 53 shapes, accurate identification requires either an accurate 3D shape percept or a KDE-alternative computation based on simultaneous measurements of 2D velocity in six positions of the display. Performance in the shape identification task improved with increased numerosity in a multidot display and with an increase in the amount of depth portrayed. Shape identification was not mediated by incidental texture-density cues but rather by motion cues derived from optic flow. The objective shape identification task is proposed as a sensitive measure of the critical aspect of kinetic depth performance. It is proposed that the structure-from-motion algorithm used by subjects to solve the KDE shape identification task involves finding local 2D velocity minima and maxima and assigning depth values to these locations in consistent proportion to their velocities.

References

Adelson, A. H. (1985). Rigid objects that appear highly non-rigid. *Investigative Ophthalmology and Visual Science, 26* (Suppl.), 56.

Andersen, G. J., & Braunstein, M. L. (1983). Dynamic occlusion in the perception of rotation in depth. *Perception & Psychophysics, 34*, 356–362.

Bennett, B. M., & Hoffman, D. D. (1985). The computation of structure from fixed-axis motion: Nonrigid structures. *Biological Cybernetics, 51*, 293–300.

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research, 14*, 519–529.

Braunstein, M. L. (1962). Depth perception in rotating dot patterns: Effects of numerosity and perspective. *Journal of Experimental Psychology, 64*, 415–420.

Braunstein, M. L. (1977). Perceived direction of rotation of simulated three-dimensional patterns. *Perception & Psychophysics, 21*, 553–557.

Braunstein, M. L., & Andersen, G. J. (1981). Velocity gradients and relative depth perception. *Perception & Psychophysics, 29*, 145–155.

Braunstein, M. L., & Andersen, G. J. (1984). A counterexample to the rigidity assumption in the visual perception of structure from motion. *Perception, 13*, 213–217.

Clocksin, W. F. (1980). Perception of surface slant and edge labels from optical flow: A computational approach. *Perception, 9*, 253–269.

Dosher, B. A., Landy, M. S., & Sperling, G. (1989). Ratings of kinetic depth in multidot displays. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 816–825.

Dosher, B. A., Landy, M. S., & Sperling, G. (in press). Kinetic depth effect and optic flow: I. 3D shape from Fourier motion. *Vision Research.*

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

Gillam, B. (1972). Perceived common rotary motion of ambiguous stimuli as a criterion of perceptual grouping. *Perception & Psychophysics, 11*, 99–101.

Gillam, B. (1976). Grouping of multiple ambiguous contours: Towards an understanding of surface perception. *Perception, 5*, 203–209.

Green, B. F., Jr. (1961). Figure coherence in the kinetic depth effect. *Journal of Experimental Psychology, 62*, 272–282.

Helmholtz, H. v. (1924). *Helmholtz's Treatise on Physiological Optics* (J. P. C. Southall, Ed. and Trans.) Optical Society of America. Reprinted by Dover Publications, New York. (Original work published 1910.)

Hildreth, E. C., & Grzywacz, N. M. (1986). The incremental recovery of structure from motion: Position vs. velocity based formulations. *Proceedings of the Workshop on Motion: Representation and Analysis, IEEE Computer Society #696, Charleston, South Carolina,*

May 7–9, 1986 (pp. 137–144). Los Angeles: IEEE Computer Society Press.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *Journal of the Optical Society of America, 72*, 888–892.

Hoffman, D. D., & Bennett, B. M. (1985). Inferring the relative three-dimensional positions of two moving points. *Journal of the Optical Society of America A, 2*, 350–353.

Hoffman, D. D., & Flinchbaugh, B. E. (1982). The interpretation of biological motion. *Biological Cybernetics, 42*, 195–204.

Koenderink, J. J., & van Doorn, A. J. (1986). Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America A, 3*, 242–249.

Landy, M. S. (1987). A parallel model of the kinetic depth effect using local computations. *Journal of the Optical Society of America A, 4*, 864–876.

Landy, M. S., Dosher, B. A., & Sperling, G. (1985). Assessing kinetic depth in multi-dot displays. *Bulletin of the Psychonomic Society, 19*, 23.

Landy, M. S., Sperling, G., Dosher, B. A., & Perkins, M. E. (1987). Structure from what kinds of motion? *Investigative Ophthalmology and Visual Science, 28*(Suppl), 233.

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

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

McKee, S. P., Silverman, G. H., & Nakayama, K. (1986). Precise velocity discrimination despite random variations in temporal frequency and contrast. *Vision Research, 26*, 609–619.

Petersik, J. T. (1979). Three-dimensional object constancy: Coherence of a simulated rotating sphere in noise. *Perception & Psychophysics, 25*, 328–335.

Petersik, J. T. (1980). The effects of spatial and temporal factors on the perception of stroboscopic rotation simulations. *Perception, 9*, 271–283.

Proffitt, D. R., Bertenthal, B. I., & Roberts, R. J. (1984). The role of occlusion in reducing multistability in moving point-light displays. *Perception & Psychophysics, 36*, 315–323.

Rogers, B., & Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception, 8*, 125–134.

Schwartz, B. J., & Sperling, G. (1983). Luminance controls the perceived 3-D structure of dynamic 2-D displays. *Bulletin of the Psychonomic Society, 21*, 456–458.

Sperling, G. (1981). Mathematical models of binocular vision. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology.* Providence, Rhode Island: *Society of Industrial and Applied Mathematics–American Mathematical Association (SIAM–AMS) Proceedings, 13*, 281–300.

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). The analysis of three-dimensional structure from moving images. In D. Ingle, M. Jeannerod, & D. Lee (Eds.), *Brain mechanisms and spatial vision* (pp. 73–93). The Hague, The Netherlands: Martinus Nijhoff.

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

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion. *Perception, 13*, 255–274.

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

Webb, J. A., & Aggarwal, J. K. (1981). Visually interpreting the motion of objects in space. *Computer, 14*, 40–49.

Williams, D., & Phillips, G. (1987). Cooperative phenomena in the perception of motion direction. *Journal of the Optical Society of America A, 4*, 878–885.

Received July 8, 1988
Revision received October 13, 1988
Accepted October 14, 1988 ■

Correction to Driver and Baylis

In the article, "Movement and Visual Attention: The Spotlight Metaphor Breaks Down," by Jon Driver and Gordon C. Baylis (*Journal of Experimental Psychology: Human Perception and Performance*, 1989, Vol. 15, No. 3, 448–456), the display durations were incorrect and should be doubled to give the correct figures. Each display frame actually lasted 40 ms. Thus, total display duration was 200 ms in Experiments 1, 3, and 4 and was 120 ms in Experiment 2.