MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1141
C.B.I.P. Memo No. 20

August, 1989

The Perceptual Buildup of Three–Dimensional Structure from Motion

Ellen C. Hildreth, Norberto M. Grzywacz, Edward H. Adelson and Victor K. Inada

Abstract: We present a set of psychophysical experiments that measure the accuracy of perceived three–dimensional structure derived from relative motion in the changing two–dimensional image. The experiments are motivated in part by a computational model proposed by Ullman (1984), called the *incremental rigidity scheme*, in which an accurate 3–D structure is built up incrementally, by considering images of moving objects over an extended time period. Our main conclusions are: first, the human visual system can derive an accurate model of the relative depths of moving points, even in the presence of noise in their image positions; second, the accuracy of the 3–D model improves with time, eventually reaching a plateau; and third, the 3–D structure currently perceived appears to depend on previous 3–D models. Through computer simulations, we relate the results of our psychophysical experiments with the predictions of Ullman's model.

© Massachusetts Institute of Technology (1989)

Acknowledgments. This report describes research done within the Artificial Intelligence Laboratory and the Center for Biological Information Processing (Whitaker College) at the Massachusetts Institute of Technology. Support for the A.I. Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124. Support for this research is also provided by the Alfred P. Sloan Foundation, the Office of Naval Research, Cognitive and Neural Systems Division, the National Science Foundation and the McDonnell Foundation.

INTRODUCTION

A valuable source of three–dimensional (3–D) information is provided by the relative motions of elements in the changing two–dimensional (2–D) image. The remarkable ability of the human visual system to recover 3–D structure from motion was explored in many early perceptual studies (for example, Wallach & O'Connell, 1953; Gibson & Gibson, 1957; White & Mueser, 1960; Green, 1961; Braunstein, 1976; Johansson, 1973, 1978; Rogers & Graham, 1979; Ullman, 1979). These studies reveal that the human system can recover the structure of rigid and nonrigid objects, under perspective and orthographic projection, and in the absence of all other cues to 3–D structure. Early perceptual work typically focused on the recovery of qualitative aspects of an object's structure, such as its apparent rigidity, volume or coherence. More recently, studies have addressed quantitatively, the accuracy of perceived 3–D structure (for example, Todd, 1982, 1984, 1985; Lappin & Fuqua, 1983; Braunstein et al., 1987; Loomis & Eby, 1988, 1989; Dosher et al., 1989; Sperling et al., 1989).

This paper presents a set of psychophysical experiments that examine both the accuracy of the 3–D model computed by the human visual system and the time course of the buildup of perceived structure. The experiments are motivated in part by a computational model proposed by Ullman (1984), called the *incremental rigidity scheme*, in which an accurate 3–D structure is built up incrementally, by considering images of moving objects over an extended time period. If this model captures some aspects of the human recovery of structure from motion, then two critical predictions arise. First, the accuracy of perceived structure should increase over an extended time period, and second, the current perception of structure should strongly influence later perceptions. Also, in contrast to previous structure–from–motion models, the incremental rigidity scheme exhibits good performance in the presence of noise in the image motion measurements, due in part to the integration of motion information over time. This observation raises the question of how well the human visual system performs in the presence of image noise. Through psychophysical experiments, we examine these three questions. Through computer simulations, we then relate the results of our psychophysical experiments with the predictions of Ullman's model.

The next section briefly reviews previous computational and perceptual studies of the recovery of structure from motion and introduces the model proposed by Ullman (1984). We then present a series of experiments in which subjects perform the simple task of ordering a set of moving points in depth. Through these experiments, we examine the accuracy of perceived structure, the nature of its buildup over time and its sensitivity to noise in the visual image. We conclude that when viewing displays containing as few as three points undergoing relative motion, the human visual system can derive quite an accurate model of the relative depths of the points, even in the presence of noise. The accuracy of the 3–D model improves with time; some observers show continued improvement up to about one second of viewing. Performance eventu-

ally reaches a plateau, beyond which there is no further improvement. After presenting the experimental results, we describe a set of computer simulations that reveal that the early time course of the buildup of perceived 3–D structure is similar to that predicted by one formulation of Ullman's model. These experiments also provide some evidence that the 3–D structure currently perceived depends on previous 3–D models. The implications of our observations for the computation of structure from motion in the human visual system are addressed in our final discussion.

THE COMPUTATION OF 3–D STRUCTURE FROM MOTION

In studying the computation of structure from motion, one immediately faces the problem that the recovery of structure is underconstrained; there are infinitely many 3–D structures consistent with a given pattern of motion in the changing 2–D image. Additional constraint is required to establish a unique interpretation. Early perceptual studies suggest that the presumed *rigidity* of objects may play a key role in the recovery of structure from motion (Wallach & O'Connell, 1953; Gibson & Gibson, 1957; Green, 1961; Jansson & Johansson, 1973; Johansson, 1973, 1977). Computational studies establish that rigidity is a sufficiently powerful constraint to derive a unique interpretation of structure under a variety of viewing conditions.

From theoretical studies, it can be concluded that by exploiting a rigidity constraint, a unique 3–D structure can be recovered from motion information alone, using image measurements that are integrated over a small extent in space and in time (for example, Ullman, 1979, 1983; Clocksin, 1980; Longuet–Higgins & Prazdny, 1980; Tsai & Huang, 1981; Hoffman, 1982; Prazdny, 1983; Kanatani, 1985; Waxman & Ullman, 1985; Mitiche, 1986; Waxman & Wohn, 1988). A review of many of these results can be found in Ullman (1983), Barron (1984) and Hildreth and Koch (1987). Theoretical studies have also given rise to algorithms for deriving the rigid 3–D structure of moving objects. Experimentation with these algorithms has revealed two important limitations. First, although it is possible in theory to recover structure from motion information that is integrated over a small extent in space and time, such a strategy may not be robust in practice (an error analysis for the case of limited temporal extent can, for example, be found in Weng, Huang and Ahuja (1989)). A small amount of error in the image measurements can lead to very different solutions (Ullman, 1983). Second, most previous algorithms derive a 3–D structure only when a rigid interpretation is possible, and otherwise do not yield any interpretation of structure or yield a solution that is incorrect or unstable.

Theoretical studies suggest that a robust algorithm for recovering structure should use motion information that is more extended in space or time. This conclusion is supported in recent computational studies (for example, Bruss & Horn, 1983; Lawton, 1983; Ullman, 1984; Adiv, 1985; Bolles & Baker, 1985; Negahdaripour & Horn, 1985; Yasumoto & Medioni, 1985; Bharwani et al., 1986; Shariat & Price, 1986; Subbarao,

1986; Landy, 1987; Waxman & Wohn, 1988; Bhanu & Burger, 1988).

With regard to the human visual system, the dependence of perceived structure on the spatial and temporal extent of the viewed motion has not yet been studied systematically, but the following informal observations have been made. Regarding spatial extent, two or three points undergoing relative motion are sufficient to elicit a perception of 3–D structure (Borjesson & von Hofsten, 1973; Lappin & Fuqua, 1983; Braunstein et al., 1987; Petersik, 1987), although theoretically the recovery of structure is less constrained for two points in motion, and perceptually the sensation of structure is weaker. An increase in the number of moving elements in view can yield a more compelling sense of 3–D shape (Todd et al., 1988; Dosher et al., 1989; Sperling et al., 1989), but its influence on the accuracy of perceived structure is unclear (see, for example, Petersik, 1980; Braunstein et al., 1987). Regarding the temporal extent of viewed motion, Johansson (1973) showed that a brief observation of patterns of moving lights generated by human figures moving in the dark (commonly referred to as biological motion displays) can lead to a perception of the 3–D motion and structure of the figures. Other perceptual studies indicate that the human visual system requires an extended time period to reach an accurate perception of 3–D structure (Wallach & O'Connell, 1953; White & Mueser, 1960; Braunstein & Andersen, 1984b; Doner, Lappin & Perfetto, 1984; Braunstein et al., 1987; Siegel & Andersen, 1988; Husain, Treue & Andersen, 1989). A brief observation of a moving pattern sometimes yields an impression of structure that is *flatter* than the true structure of the moving object. Thus, the human visual system can derive some sense of structure from motion information that is integrated over a small extent in space and time. An accurate perception of structure, however, may require a more extended viewing period.

The sensitivity of early structure–from–motion algorithms to error in the image motion measurements raises the question of how sensitive is the human recovery of structure to image noise. Lappin, Doner and Kottas (1980) showed that small amounts of noise could disrupt subjects' ability to discriminate between different amounts of coherence in structure–from–motion displays. This study used only two frames in alternation, however. Other experiments have shown that subjects can tolerate larger amounts of noise when extended sequences of images are used (Petersik, 1979; Doner, Lappin & Perfetto, 1984; Todd, 1984, 1985; Husain, Treue & Andersen, 1989). Todd's studies, in particular, show that subjects can make an accurate assessment of 3–D shape and motion in the presence of large amounts of visual noise.

Although most algorithms for recovering structure from motion are unable to interpret nonrigid motions, there are exceptions to this that can interpret restricted classes of nonrigid motions (for example, Rashid, 1980; Hoffman & Flinchbaugh, 1982; Bennett & Hoffman, 1985; Koenderink & Van Doorn, 1986; Subbarao, 1986). The mechanism for recovering structure from motion in the human visual system appears not to be based strictly on the rigidity assumption. It is an everyday experience to

perceive the structure and motion of deforming objects such as a flowing river, an expanding balloon, or a dancing ballerina. Such experiences are rich with many cues to 3–D structure. In controlled perceptual studies that isolate relative movement as a single cue to 3–D structure, however, it also appears that the human visual system can derive some sense of structure for a broad range of nonrigid motions, including stretching, bending and even more complex types of deformations (for example, Johansson, 1973, 1978; Jansson & Johansson, 1973; Cutting, 1982; Todd, 1982, 1984, 1985; Loomis & Eby, 1988, 1989). Furthermore, displays of rigid objects in motion sometimes give rise to the perception of somewhat distorting objects (Wallach, Weisz & Adams, 1956; White & Mueser, 1960; Braunstein, 1976; Sperling et al., 1989; Schwartz & Sperling, 1983; Braunstein & Andersen, 1984a; Adelson, 1985; Loomis & Eby, 1988, 1989).

Recently, Ullman (1984) proposed a more flexible method for deriving the structure of rigid and nonrigid objects that provides a natural means for integrating motion information over an extended time period. This method makes use of the rigidity assumption, but in a more flexible way than previous studies. The algorithm, called the *incremental rigidity scheme*, maintains an internal model of the structure of a moving object, which is continually updated as new positions of image elements are considered. The initial model may be flat, if no other cues to 3–D structure are present, or it may be determined by other cues available, for example, from binocular stereopsis, shading, texture or perspective. As each new view of the moving object appears, the algorithm computes new 3–D coordinates for points on the object, which maximize the rigidity in the tranformation from the current model to the new positions. In particular, the algorithm minimizes the change in the 3–D distances between points in the model. The formulation presented by Ullman assumes the input to the recovery process to consist of a sequence of discrete frames, each containing a set of discrete feature points whose positions are obtained by orthographic projection of the scene onto the image plane. Through the process of repeatedly considering a new frame in the sequence and updating the current model of the structure of the moving features, the incremental rigidity scheme builds up and maintains a 3–D model, and can be applied both to rigid and nonrigid objects in motion. Recent extensions to the incremental rigidity scheme use velocity information directly as input to the recovery of structure from motion, and perspective projection (Grzywacz & Hildreth, 1987; for further details, see also Grzywacz & Hildreth, 1985). Landy (1987) presents a parallel structure–from–motion model that implements a similar scheme in a cooperative network. Details of three different formulations of the incremental rigidity scheme addressed in this paper appear in the section on computer simulations. Other models have been suggested that impose rigidity by requiring that the 3–D distances between points in space change very little from one moment to the next (for example, Mitiche, 1986; Subbarao, 1986; Jasinschi & Yuille, 1989), although these models do not build up 3–D structure incrementally as in Ullman's proposed scheme.

The features of the incremental rigidity scheme that distinguish it from other models are the buildup of an accurate 3–D model over an extended time and the use of a current 3–D model as an explicit source of constraint on the model computed at the next moment. In most other structure–from–motion models, the computed 3–D structure at each moment is constrained only by direct visual input that is integrated over a small window in time. The next sections explore whether the recovery of structure from motion in the human visual system exhibits these salient properties.

GENERAL METHODS

We first describe aspects of the visual stimuli and experimental procedure that are common to the entire set of experiments. The experimental design was guided by a number of considerations. First, the task relies on an objective judgement that does not require the observer to form an internal subjective scale of properties such as 3–D distance or amount of rigidity. Second, the subject is able to perform the task in a sufficiently short time period that we can measure the time course of the early buildup of 3–D structure. If, on the other hand, it only required one or two seconds for the visual system to compute an accurate model of 3–D structure, but several seconds of observation time were needed for the subject to make a judgement, then we could only assess the accuracy of the final computed 3–D model. Such a task would not allow us to explore the intermediate structures perceived in the first one or two seconds of viewing time. Third, we have designed a task that relies as much as possible on the derived 3–D structure of the moving elements, rather than their raw 2–D positions or velocities in the display. Because the recovery of 3–D structure from motion necessarily relies on properties of the changing 2–D projection, one cannot guarantee that observer's judgements are not based directly on 2–D information, but we chose an experimental design that makes it very difficult for the observer to use 2–D cues directly. Finally, the experiments make use of a quantitative judgement that requires only relative movement as the source of 3–D information.

Subjects

The authors, who are all trained psychophysical observers, served as the subjects for these experiments.

The Visual Stimuli

A set of three points distributed in space was rotated around a central axis and projected onto a 2–D computer display, using orthographic projection. To describe the stimuli in more detail, let us assume a coordinate system in which the X and Y axes are the horizontal and vertical axes in space and in the image plane, which are the same under orthographic projection, and the positive Z axis is directed perpendicular

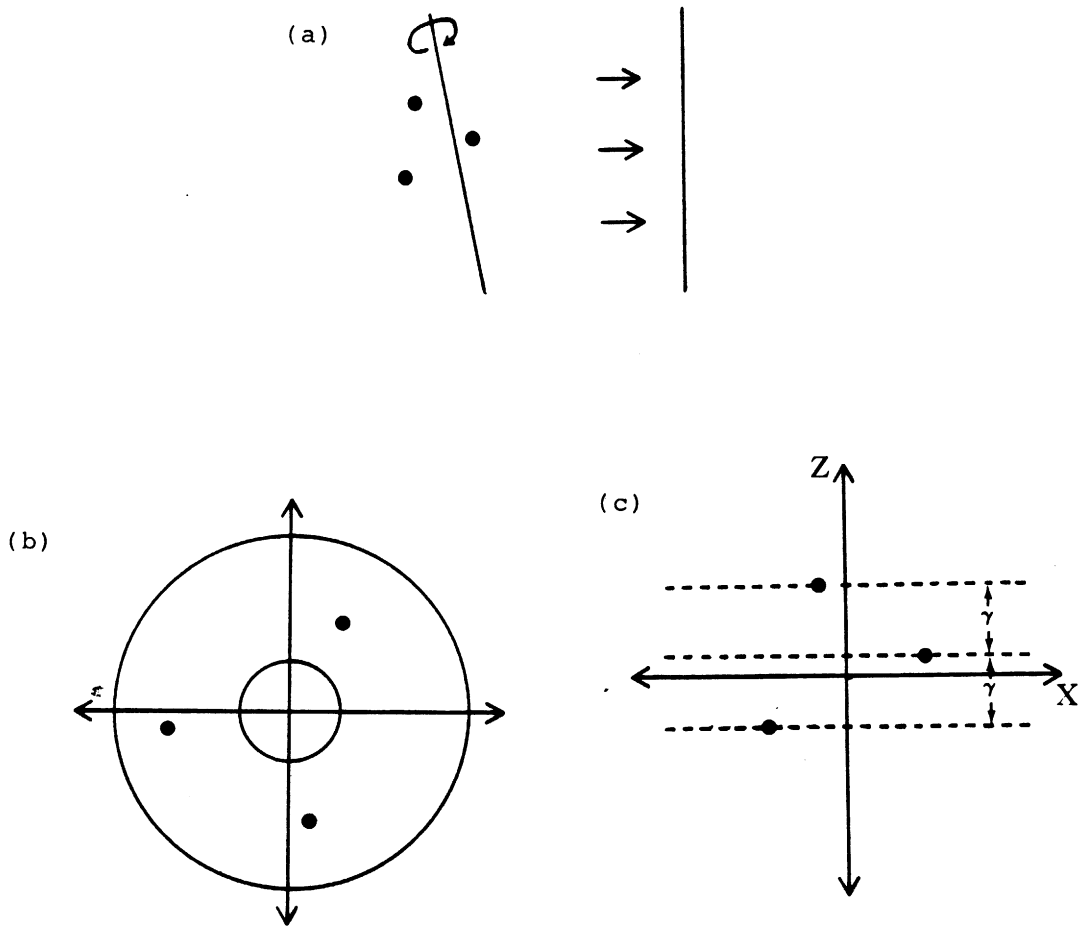

Figure 1. The ordinal experiment. (a) Side view of the experimental setup, indicating the slanted axis of rotation. (b) Projection of a typical configuration of points onto a plane perpendicular to the axis of rotation. The circular outlines indicate the annulus within which the points are located. (c) Bird's eye view of the final positions of the three points, indicating their separation in depth, γ.

to the image plane, away from the viewer. For the first experiment, the axis of rotation of the three points was slanted 10° away from the image plane, as shown in Figure 1a. For later experiments, the axis of rotation was parallel to the image plane. When projected onto a plane perpendicular to the axis of rotation, the positions of the three points always lie within an annulus, as indicated in Figure 1b. The outer boundary of the annulus restricts the overall range of X and Z coordinates of the moving points. The reason for restricting the points to lie outside the inner boundary of the annulus is that points located near the center of this circular projection would move very little under rotation. A point that is moving by only a small amount could easily be identified as lying near the center of the cylindrical volume encompassing the set of three points. The use of an annulus as shown in Figure 1b removes this potential cue

to 3–D structure.

The particular configurations of points were chosen such that for any given rotation, the positions of the points were evenly spaced in depth in the final frame that was viewed. Let γ denote the displacements in depth for this final view, as shown in Figure 1c. The parameter, γ, is defined in units of picture elements on the display. In the plane of the display, which is perpendicular to the observer's line of sight, one picture element corresponds to a visual angle of $2.5'$.

The Y coordinates of the three points (vertical positions in the image plane) were chosen such that there was always a minimum separation between vertical positions of $25'$ of visual arc. The size of each point was $10' \times 10'$ of visual angle and the resolution of the display was such that the positions of the points could be set to a resolution of $2.5'$. The overall size of the window of the computer display in which the points appeared was $10° \times 10°$.

The display itself was a monochrome video monitor from a Symbolics Lisp Machine, with a fast decaying (P4) phosphor. The experiments used black dots on a white background, in order to reduce the possible effects of persistence of the display.

For each trial in the experiment, the points were rotated through a given total angular extent in increments of $1.5°$ around the central axis. A discrete frame was created for each angular position of the points and the entire set of frames was displayed as a movie. The presentation time for each frame was approximately 33 msec. There was no interstimulus interval (ISI) — after the 33 msec presentation time, each frame was immediately replaced by the next frame in the sequence. A fixation mark also appeared in every frame and the subject was required to fixate on the mark throughout the duration of each movie.

Experimental Procedure

For each trial, the first frame of the movie appeared on the display, the subject pushed a button to indicate that he was ready; the movie was then displayed. The distance of the viewer from the display was 0.4 meters. Viewing was monocular and in a dark room.

The subjects were asked to specify which of the three points was located midway in depth between the other two. Because orthographic projection is used, there are two possible rigid structures corresponding to the changing projections — one is the structure used to generate the frames and the second is its reversal in depth. Both solutions share the same central point, so the outcome of a given trial is not effected by whether the subject sees a given structure or its depth reversal.

To specify their choice to the computer, subjects were given a box with three buttons oriented vertically and were told to associate the vertical ordering of the three buttons on the box with the vertical ordering of the positions of the points on the display. After each trial, the subject pushed the appropriate button depending on

whether the top, middle or bottom point was perceived as being between the other two points in depth. No feedback regarding the correctness of the response was given. This lack of feedback reduced the likelihood for subjects to use simple tricks based directly on 2–D cues for performing the task.

The accuracy of perceived 3–D structure can be assessed by measuring how well subjects perform this task as we vary the separation between the points in depth, γ, while the time course of the buildup of this accuracy can be assessed by measuring subjects' performance as the angular extent of rotation of the points is increased. The particular stimulus parameters used are indicated in the discussion of each individual experiment.

EXPERIMENT 1

This first experiment addresses the time course of the buildup of accuracy of perceived 3–D structure, by measuring subjects' performance as the angular extent of rotation is varied.

The Visual Stimuli and Experimental Procedure

Six different angular extents were used in this experiment: $0°, 3°, 6°, 15°, 30°, 45°$. A single experimental session consisted of 324 trials with 81 different configurations of the three points. For this first experiment, the displacement in depth was $\gamma = 70$. Through a single session, the 81 configurations were repeated four times in random order. The different angular extents appeared in blocks of 27 trials, with the ordering of the blocks randomized. The $0°$ condition served as a control, to show that there were no static cues to depth in these displays. For each experimental condition, we computed the percentage of correct responses by the subject.

The axis of rotation was slanted $10°$ away from the image plane, as shown in Figure 1a. The use of the slanted rotation axis made it possible to design the 81 configurations in a way that made it very difficult for subjects to base their judgement of relative depth directly on 2–D information about positions or velocities in the image. There was no bias for particular positions on the computer display, that is, the point corresponding to the correct response was equally likely to appear on the top, middle or bottom in the vertical direction, and to the left, right or middle in the horizontal direction. Also, there was no bias for particular image velocities, that is, the correct point was equally likely to move with the highest, middle or lowest image velocity. Consequently, it was not possible for the subjects to base their judgement directly on simple properties of the 2–D image positions or velocities of the points. The use of the slanted rotation axis was essential for removing the 2–D velocity cue. If the axis of rotation were parallel to the image plane, then projected image velocity could provide a direct cue to depth.

Experimental Results

For this first experiment, data was gathered for two subjects, ECH and NMG, and is displayed in Figure 2. The angular extents of rotation are indicated on the vertical axis of the graph shown in Figure 2 and the percentage of correct responses appears on the horizontal axis. Each data point represents the result of 324 trials. Error bars indicate a single standard error of the mean. A chance level of performance corresponds to 33%. Both subjects performed at chance for 0° of rotation.

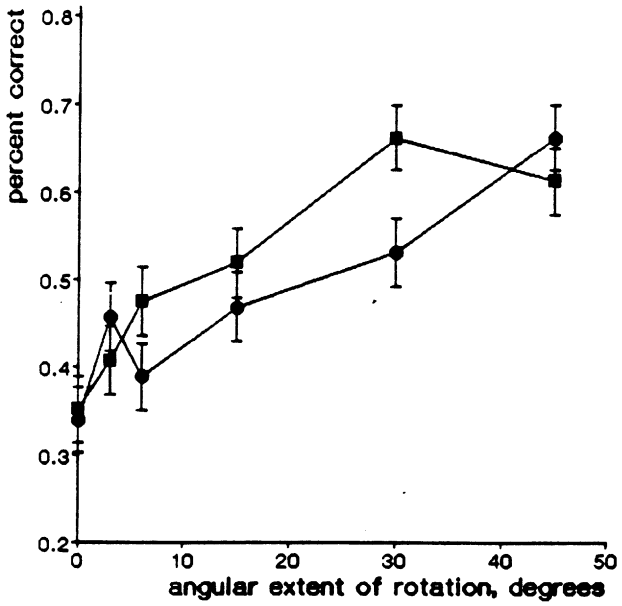

Figure 2. Results of Experiment 1. Data for ECH (squares) and NMG (circles) are shown superimposed. The percentage of correct responses is plotted against the total angular rotation. Standard errors are displayed as vertical bars. The data shows improvement in performance with increased angular extents of motion.

The first observation that can be made is that subjects did exhibit a buildup in the accuracy of the perceived 3–D structure of the points. There was a rise in performance level from 0° to 30° of rotation for the subject ECH (squares), while performance for the subject NMG (circles) continued to rise up to 45°. Later experiments indicate that performance typically reaches a plateau at 30° and 45° of rotation for the subjects ECH and NMG, respectively. For the longer angular rotations, subjects found it increasingly difficult to maintain the perception of a rigid structure over the full extent of rotation; the points sometimes appeared to move independently of one another in a nonrigid manner.

It should be emphasized that the total viewing time and total angular extent of viewed motion are directly coupled in these visual displays. We cannot conclude

from this experiment alone whether the accuracy of perceived structure depends most critically on the extension of the visual stimulus in time, across space, or both. If the extent in time alone is a factor in determining this accuracy, we note that the rotation of 30° corresponds to a total viewing time of 768 msec, which is in rough agreement with the time expected from the studies of Andersen and Siegel (1988; Siegel & Andersen, 1988; see also, Husain, Treue & Andersen, 1989) indicating that several hundred msec of viewing time is required to make visual discriminations of motion that are essential to the detection of 3–D structure.

EXPERIMENT 2

The second experiment explores both the temporal buildup of perceived 3–D structure and the accuracy of the perceived structure. Accuracy is assessed by measuring subjects' performance as the final separation in depth between the points, γ, is varied.

The Visual Stimuli and Experimental Procedure

The experimental setup is similar to that of the first experiment, but with an important exception. The points were now rotated around a central vertical axis (that is, parallel to the image plane). The reason for this was a practical one. In the first experiment, there was no bias for particular image velocities, in that the point that was midway in depth at the end of the rotation was equally likely to move with the slowest, middle or fastest velocity in the image. Because of the nature of the geometric projection, for smaller final separations of the points, γ, this uniform distribution of velocities could only be achieved if the points were allowed to have large vertical separations. The task itself, however, becomes difficult when the points are widely separated in the image, because the points appear to decouple from one another. We therefore decided to use a vertical axis of rotation in this experiment, allowing the configurations to become more compact in the vertical direction. While this introduces a potential 2–D velocity cue, we believe that the subjects were not using this velocity cue directly, for reasons that are elaborated upon in our discussion of the results of this experiment.

A single experimental session consisted of 256 trials and used 64 different configurations of the three points and 4 different angular extents of rotation. The displacement in depth, γ, was kept constant for a single session, but was now varied between sessions. Through a single session, the 64 configurations were repeated four times in random order. The different angular extents appeared in blocks of 16 trials, with the ordering of the blocks randomized. As a control, the points again remained stationary in some trials. The 64 configurations were chosen such that there was no bias for particular positions on the computer display (the point corresponding to the correct response was equally likely to appear on the top, middle or bottom in the vertical direction, and to the left, right or middle in the horizontal direction). A total of 8 different angular

extents were used in this experiment: $0°, 6°, 15°, 30°, 45°, 60°, 90°, 177°$. In addition, 5 different depth displacements, γ were used: $10, 20, 30, 40, 50$. For each experimental condition, we computed the percentage of correct responses by the subject.

Experimental Results

Complete data were gathered for three subjects, ECH, VKI and NMG. Partial data was also obtained for a fourth subject, which confirmed the general trends seen in the data. Individual data for ECH and NMG are displayed in Figures 3a and 3b, respectively. The data are shown separately, because quantitative differences between the performance of individual subjects were observed. The two sets of data shown here represent the range of performance observed. Figure 3c shows the results of averaging the data obtained for the three subjects. The angular extents of rotation are indicated on the horizontal axis of the graphs shown in Figure 3 and the percentage of correct responses appears on the vertical axis. Data for different displacements γ are drawn on separate curves. As a control, we again verified that subjects were at a chance level of performance for 0° of rotation. Each data point in Figure 3 represents the result of 64 trials (that is, 64 different configurations).

It can again be observed that subjects showed a buildup in the accuracy of the perceived 3–D structure of the points. Particularly for the larger values of γ, there was a steady rise in performance level from 0° of rotation that reached a plateau at about 30° of rotation for subject ECH (Figure 3a) and about 45° of rotation for NMG (Figure 3b). After only 6° of rotation, subjects already reached a level of performance that was often within about 15% of the level at which performance reached a plateau. After about $30° - 45°$ of rotation, performance generally did not continue to improve with larger extents of motion. The level of performance dropped for very large angular extents. (When tested against a binomial distribution with probability 0.5 that the level of performance was larger at 180° than at 90°, a significant drop was found ($n = 18, \quad P < 0.05$).) Subjects again found that it was difficult to maintain the perception of a rigid structure over such a long viewing period; the points sometimes appeared to move independently of one another in a nonrigid manner.

Performance was generally worse for smaller γ. For the largest value of γ tested, $\gamma = 50$, performance for some subjects reached the 90% level of performance. For the smallest displacement, $\gamma = 10$, subjects were not significantly above chance, while for $\gamma = 20$, subjects were well above chance for all angular extents of rotation. This suggests that the threshold for discrimination of relative depth from motion may lie somewhere in this range of displacements. For the case where $\gamma = 20$, the relative depth between the points is 1.45% of the distance between the observer and computer display, that is, the threshold was roughly 0.6 cm from a viewing distance of 40 cm. Measurements need to be made at additional viewing distances, however, to derive a reliable measure of structure–from–motion *acuity*.

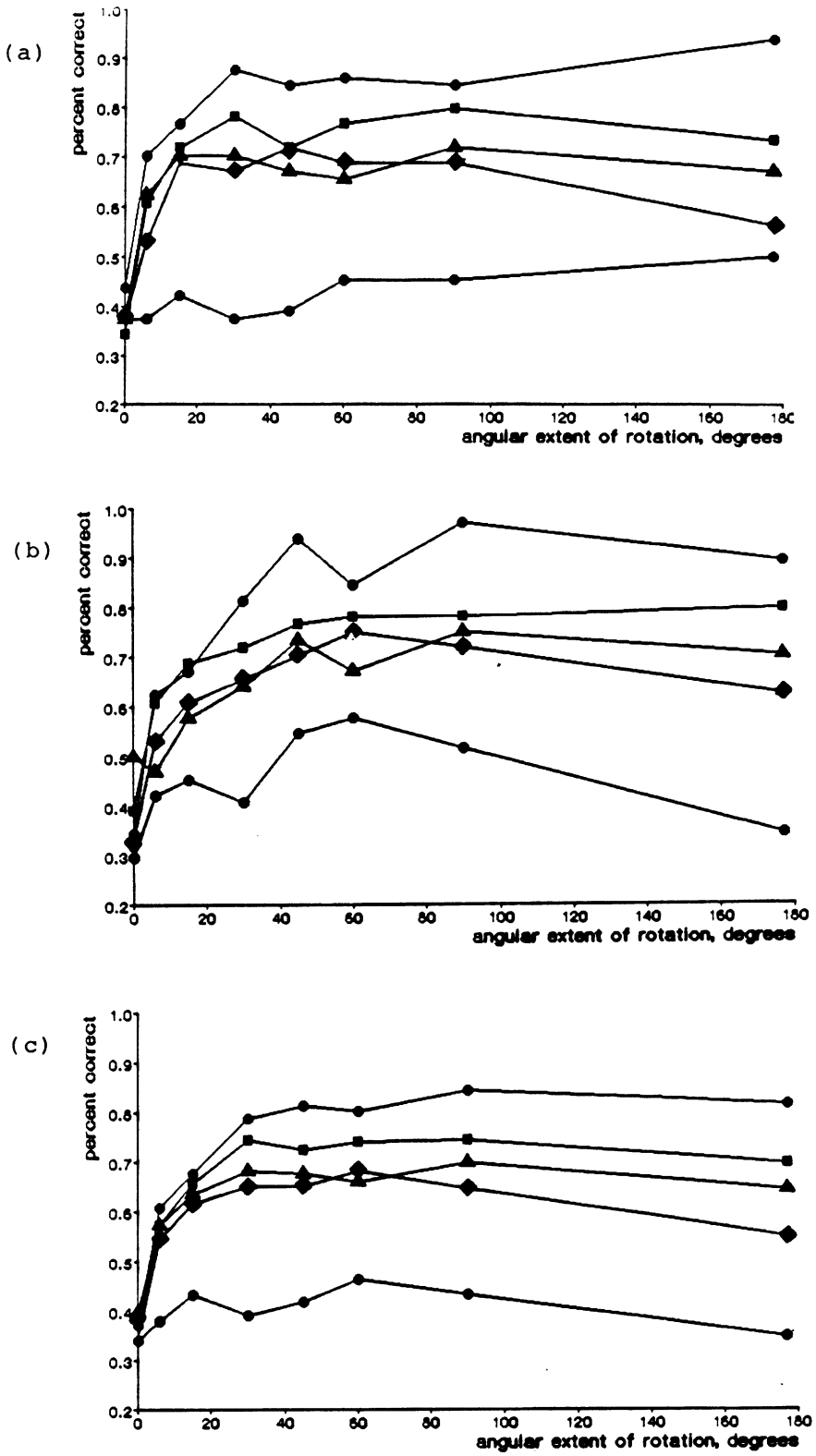

Figure 3. Results of Experiment 2. The percentage of correct responses is plotted against the total angular rotation. The five different curves in each figure correspond to the five displacements in depth, $\gamma = 10$ (circles, lower curve), 20 (diamonds), 30 (triangles), 40 (squares) and 50 (circles, upper curve). (a) Data for subject ECH. (b) Data for subject NMG. (c) Average data for three subjects. Performance improves both with angular extent of rotation and with increased displacement in depth.

Comparing the data here to that of Experiment 1, we can see that the overall level of performance is significantly higher in this case. There are at least three possible reasons for this. First, because we are using vertical axis rotation and orthographic projection here, there is no coupling between the vertical positions of the points in the image and their position in depth. This allowed us to construct configurations that were more compact in the vertical direction. The relative 3–D structure is easier to judge when the projected points are closer to one another. Second, the accuracy of the 3–D structure derived from structure–from–motion algorithms typically degrades as the axis of rotation is slanted further away from the image plane. This occurs because, in general, the amount of relative motion between points that is due to their relative depths decreases as the axis of rotation is slanted further away from the image plane. If the rotation axis is slanted by 90°, so that it is now perpendicular to the image plane, there is no relative movement due to relative depths. Computer simulations with the incremental rigidity scheme (Hildreth and Grzywacz, unpublished observations) show a steady decline in the accuracy of computed 3–D structure as the angle of slant of the rotation axis is increased from 0° to 90°. Loomis and Eby (1988, 1989) recently showed that the human visual system also exhibits this behavior, which could contribute to a drop in performance in Experiment 1, where the axis of rotation is slanted 10° from the image plane.

A third reason for the improved performance in Experiment 2 may be the introduction of the potential 2–D velocity cue into the experimental setup. The point that is midway in depth also has a velocity that is between that of the other two points. Subjects were instructed to base their judgements on perceived 3–D positions of the points and subjectively reported doing so, but may have inadvertently used the 2–D velocity cue, at least on some trials. We believe, however, that subjects were not making sole use of the 2–D velocity cue, for the following reasons. First, when we began these experiments, we were not aware of this potential 2–D cue, and found later that if we explicitly try to use this cue, we perform substantially better. Second, psychophysical studies indicate that the size of the temporal integration window for measuring image velocities ranges between 80 msec for high velocities to about 200 msec for medium range velocities (see, for example, McKee & Welch, 1985). Integration times for measuring very slow velocities may be longer (S. McKee, personal communication), but such low velocities rarely occur in the visual stimuli used here. The buildup in accuracy of perceived 2–D image velocities therefore cannot account for the more extended temporal buildup in performance that we see here, which lasts on the order of 800 – 1000 msec. If the subjects based their judgements on 2–D velocities directly, one might expect performance to reach a plateau at only about 200 msec. Third, the consistency of the time course measured here with that measured in Experiment 1, where the direct 2–D velocity cue was not available, also suggests that the buildup of accuracy largely reflects the increased accuracy in perceived 3–D structure rather

than perceived image velocities. In particular, the experiments indicate a consistent plateau in performance level at 30° for ECH and 45° for NMG.

EXPERIMENT 3

It was noted earlier that a common limitation of many structure–from–motion algorithms is extreme sensitivity to noise in the visual image. The incremental rigidity scheme, however, is quite robust against noise, in part because it integrates visual information over an extended time period and also because it allows deviations from rigidity.

Experiment 3 examines the nature of the degradation in human performance on the same ordinal task described earlier, as a function of the amount of noise introduced in the stimulus. In the section on computer simulations, we compare the psychophysical data with the behavior of the incremental rigidity scheme.

The Visual Stimuli and Experimental Procedure

The experimental procedure used here was similar to that used in Experiment 2. The visual stimuli differed in the following way. In each discrete frame, Gaussian distributed noise was added to the X and Y positions of the points in the projected image. The space constant σ for the Gaussian was held constant throughout a single experimental session, and varied between sessions. The levels of noise used here were sufficiently large that the erratic motion of the points was very apparent.

Experimental Results

Individual data for two subjects, ECH and NMG, are shown in Figures 4a and 4b, respectively. Only a single displacement in depth, $\gamma = 40$, was used in this experiment. The subject ECH performed the experiment with added Gaussian noise for which $\sigma = 2.0$ and 4.0 (expressed in terms of visual angle, $\sigma = 5'$ and $10'$ of visual arc). The subject NMG performed the experiment with $\sigma = 2.0$, 4.0 and 6.0.

For the subject ECH (data shown in Figure 4a), the added noise uniformly degraded performance for all of the angular extents of rotation. The drop in performance for smaller rotations was not significantly greater than that for larger rotations. Subjectively, the task appeared much more difficult for the larger level of noise, but the difference in mean performance between the two noise levels was small. The effect of the noise was qualitatively similar for the subject NMG (data shown in Figure 4b), although the decrease in performance for $\sigma = 2.0$ and 4.0 was smaller than that seen in the data for the subject ECH. For each level of noise, a plateau in performance was reached after 45° or 60° of rotation. Data collected for a third subject was essentially the same as that shown in Figure 4a.

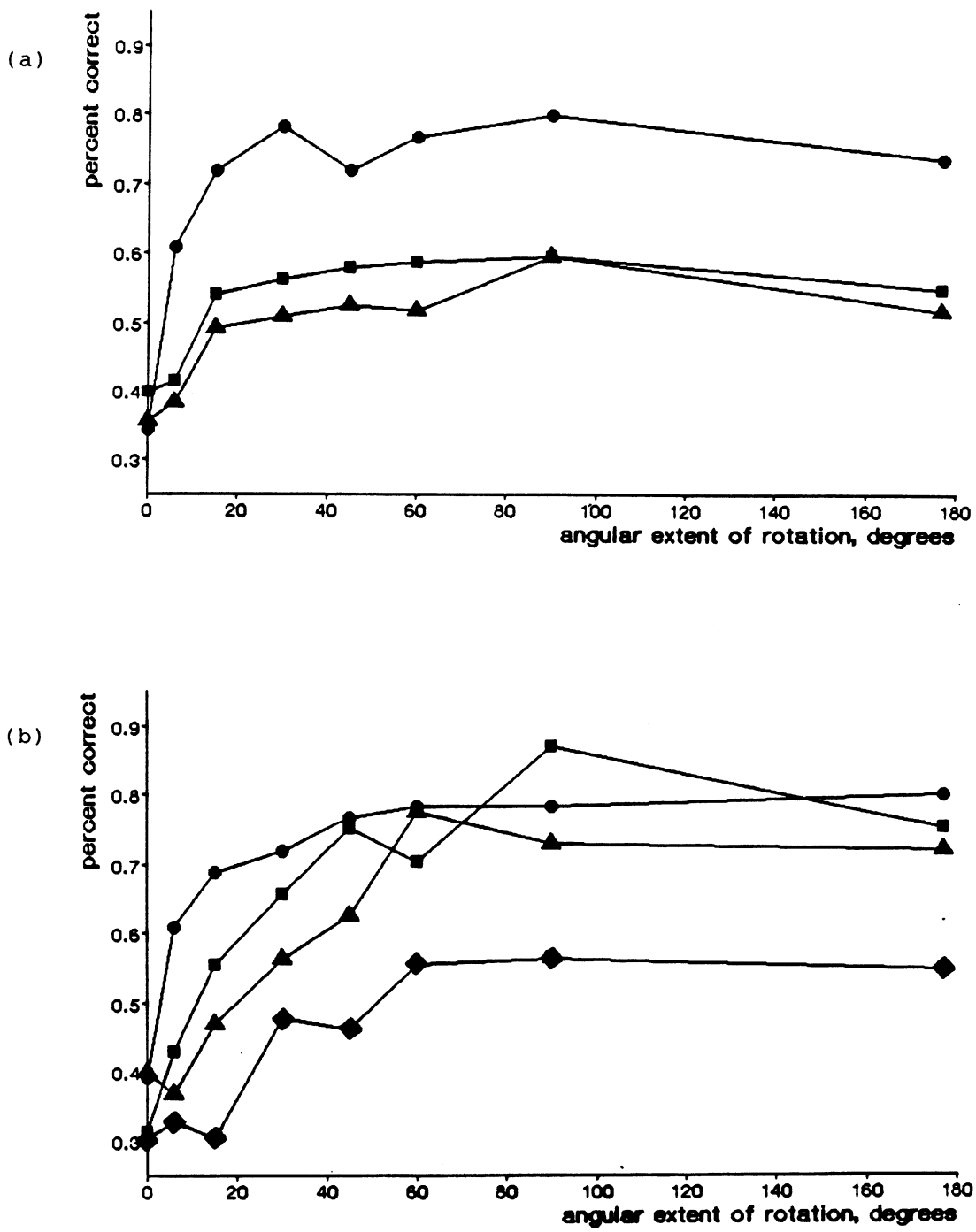

Figure 4. Results of the noise experiments. The percentage of correct responses is plotted against the total angular rotation. The different curves correspond to different levels of added Gaussian noise; $\sigma = 0.0$ (circles), $\sigma = 2.0$ (squares), $\sigma = 4.0$ (triangles) and $\sigma = 6.0$ (diamonds). $\gamma = 40$ for all experimental sessions. (a) Data for subject ECH. (b) Data for subject NMG. There is a gradual degradation in performance with increased levels of noise.

The overall data for the subject NMG indicates less sensitivity to noise than that demonstrated for the subject ECH. Recall that in the first two experiments, NMG showed a slower buildup in the accuracy of his internal model of the 3–D structure of the points. This slower buildup may suggest a longer integration time for recovering structure, which is likely to yield the lower sensitivity to noise exhibited in this experiment.

It is significant that performance at this task does not entirely break down with the large levels of noise used here, that is, does not drop quickly to a chance level of performance. The degradation for small angular extents of motion is expected, as the added noise sometimes makes the displacements of the points totally incorrect, given their true 3–D structure, for all or most of their extent of motion. Subjectively, it was observed that for the large angular extents, the "average" displacement of the points over their full trajectory can still be judged and used to interpret their rough 3–D structure. The data here suggests that the human visual system may not rely on precise measurements of the velocities and accelerations of image features, but rather may require only rough estimates of the positions or motions of image features, perhaps over an extended time period.

EXPERIMENT 4

It was noted earlier that a second salient feature of the incremental rigidity scheme is the dependence of the current 3–D model on past 3–D models. This is different from other structure–from–motion algorithms, in which the computation of 3–D structure at a particular moment in time depends only on the visual input measured over some limited time frame. This last experiment attempts to test whether perceived 3–D structure depends on previous 3–D structures derived by the visual system.

Early observations from computer simulations with the incremental rigidity scheme indicated that the algorithm sometimes behaves differently when started with different initial models of the 3–D structure of a set of points (Hildreth and Grzywacz, unpublished observations). In general, if the algorithm begins with an initial model that is compressed relative to the true structure of the object, then the algorithm will converge toward the true structure. If, on the other hand, the algorithm begins with a model that is very stretched (say, by a factor of two) relative to the true structure, then the algorithm often settles into a stretched and slightly nonrigid structure, rather than converging toward the true, rigid structure. In this experiment, relative movement was used to establish an initial perception of 3–D structure that was different from the structure on which the observer was to be tested. We examined the influence of different initial 3–D models on the subsequent perception of structure.

Much of the experimental setup here was similar to that used in Experiment 2. For each trial, however, the motion of the three–point configurations viewed in Experiment 2 was now immediately preceded by the motion of a different configuration of three

points. The last view of the preceding configuration was arranged to coincide with the first view of the configuration on which the observer was tested, so that there was no abrupt transition between the motion of the first and second configurations. The experiment examined three different types of motion preceding the movement of the test configurations:

(1) In the first condition, the preceding configurations were both stretched relative to the test configurations and the ordering of the points in depth was randomized. The ratio between the stretched and true depths was, on average, between 4:1 and 5:1. Both the stretched and test configurations were rotated around a central vertical axis.

(2) In the second condition, the preceding configuration was such that at the transition point at which the motions of the two configurations join, the preceding configuration was flat and in the plane of the computer display. Both the planar and test configurations of points were rotated around a central vertical axis.

(3) In the third condition, the initial configuration was identical to the test configuration, but was rotated around the line of sight. (This rotation should not convey any information about the real 3–D structure of the points.) After being rotated around the line of sight, the configuration was then rotated around the vertical axis as before.

Based on observations from computer simulations, we expect that if the incremental rigidity scheme is an appropriate model for the human recovery of structure from motion, then the first manipulation above should lead to a substantial degradation in the subjects' ability to judge accurately the structure of the test configuration. This is because we are initially *priming* the subject to see a stretched 3–D configuration, in which the ordering of the points may be different from that of the test configuration. Suppose that the initial stretched configuration is used by the visual system as an explicit source of constraint on the subsequent recovery of the 3–D structure of the test pattern. The test pattern might then be forced to look somewhat stretched and with an incorrect ordering of the points in depth relative to the true structure. If this were the case, one would expect a degradation in the quality of the 3–D structure attributed to the test pattern. We might not expect perceived structure to remain incorrect indefinitely, however. Internal noise, for example, might initiate changes in perceived structure. In addition, observations by Adelson (1985) suggest that compact views of a rotating 3–D object tend to be interpreted as the projection of a compact object in 3–D space, rather than a stretched object viewed from an unusual angle. Thus, an object that cycles between stretched and compact 2–D views typically appears to distort continuously. This phenomenon is also likely to occur in our displays.

The second and third manipulations, on the other hand, should not lead to sub-

stantial degradation in performance. This expectation is based on the observation that in general, the incremental rigidity scheme will converge quickly to the true structure of a rotating object, if it begins with an initial model that is compressed in depth, relative to the true structure. The algorithm assumes that in the absence of other 3–D cues, the initial model is flat and parallel to the image plane. The second and third manipulations described above only serve to strengthen the perception of a flat initial 3–D structure, prior to the motion of the test configuration.

The Visual Stimuli and Experimental Procedure

For each condition, experimental sessions were run with and without the previous configuration. In the case of condition (1), the stretched points were rotated around the central vertical axis for 36° in increments of 1°. The stretched pattern was then immediately followed by the rotating test configurations, which were constructed as described in Experiment 2. For condition (2), the flat configuration of points was rotated around the central vertical axis for 45° in increments of 1.5°, and then followed immediately by the test configuration. For condition (3), the configurations of points were initially rotated by 90° around the line of sight in increments of 1.5° before being rotated around the central vertical axis. For all three conditions, the test configurations were rotated only for small total angular extents (6°, 15°, 30° and 45° of rotation), in increments of 1.5° of rotation per frame. The subjects ran two sessions, each containing 256 trials, and the percentages of correct responses were calculated.

Experimental Results

Average data for the subjects NMG, ECH and VKI are shown in Figure 5 for condition (1), in which the initial configuration is stretched in depth. Each data point represents the results of 384 trials. We show both the data for the control experiment (circles), in which only the test configuration appeared, and the data for the case in which the stretched configuration appeared first (triangles). The vertical bars indicate standard errors.

The presence of the initial stretched configuration degrades the subsequent computation of the structure of the test configuration. The drop in performance was large for 6° and 15° of rotation, but there was no significant drop in performance for 30° and 45° of rotation. In terms of the total time over which performance was affected, individual data indicated that this influence extended for about $500-600$ msec for the subjects ECH and VKI, while for the subject NMG, an influence could still be seen after about 1200 msec.

Figures 6 and 7 show average data for the subjects ECH and NMG, for the two control conditions, in which the preceding configuration was flat and rotated around the vertical axis, or rotated around the line of sight. The data points each represent the results of 256 trials. In the case of Figure 6, there is a small, but significant, drop

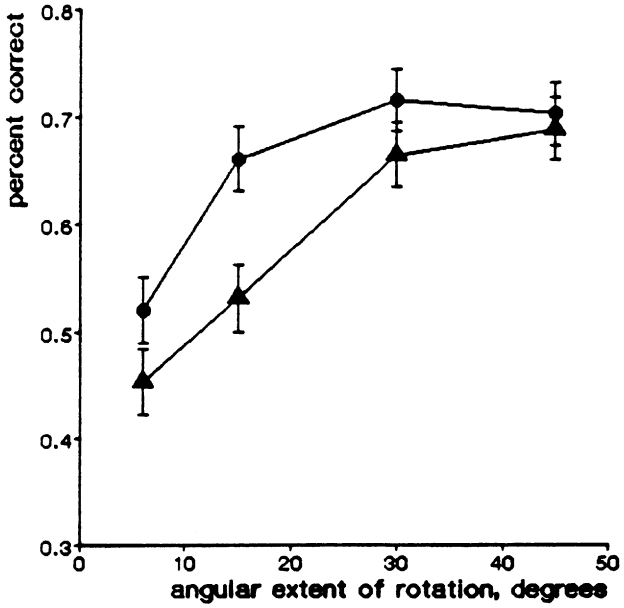

Figure 5. Average data for the experiment with an initial stretched structure, for subjects ECH, NMG and VKI. Circles correspond to the control condition and triangles to the case where the stretched configuration appeared first. Vertical bars indicate standard errors. The initial stretched configuration leads to degradation in performance that lasts up to about 30° of rotation of the test configuration.

in performance for the smallest angular rotation, but there are otherwise no significant differences between the data for the control and test conditions.

Some of the degradation in performance seen in this experiment can be attributed to the size of the temporal integration period that is used to measure retinal image motion. It appears, however, that this integration period could only account for the degradation seen for the smallest angular extent of rotation, 6°, which extended over a viewing time of roughly 150 msec. As we noted earlier, psychophysical studies indicate that the size of the temporal integration window for measuring image velocities ranges between 80 msec for high velocities to about 200 msec for medium range velocities (see, for example, McKee & Welch, 1985). We therefore do not believe that the extended influence seen here (up to a second or so) can be accounted for on the basis of the mechanisms by which retinal image motion is first measured. In further support of these conclusions, the second and third manipulations used to generate the preceding configurations (data shown in Figures 6 and 7) should also influence the performance of the motion measurement stage, but did not lead to an extended influence on the quality of the perceived 3–D structure.

For the case of the incremental rigidity scheme, the structure computed at a particular moment appears to depend on previous 3–D models. Based on computer

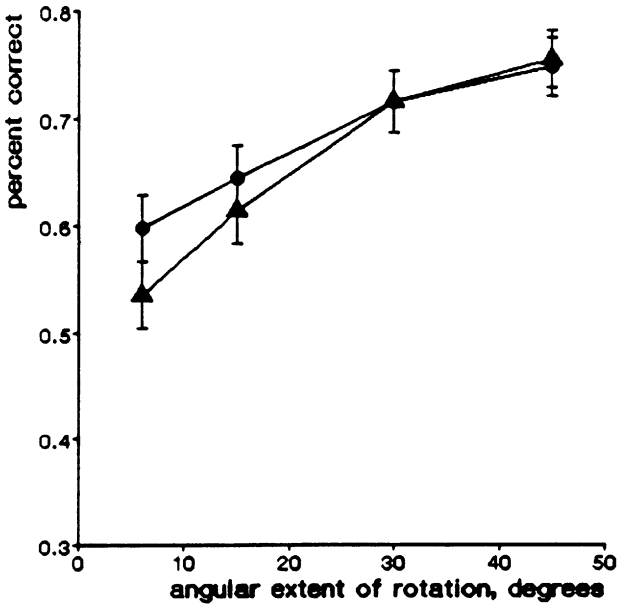

Figure 6. Average data for the experiment in which an initial planar configuration is presented, for subjects ECH and NMG. Circles correspond to the control condition and triangles to the case where the planar configuration appeared first. Vertical bars indicate standard errors. There is degradation in performance only for the smallest angular rotation, in contrast to the results obtained for the initial stretched configuration shown in Figure 5.

simulations with this model, we expected that if observers initially viewed a stretched configuration, then there would be a significant degradation in the quality of the structure perceived at later times. On the other hand, if a flat configuration were viewed initially, we expected no degradation at later times (except for the shortest angular extent of rotation). Our experimental data showed these expectations to hold true for a limited time frame of a second or so. When the initial configuration was stretched, however, the perceived 3–D structure eventually 'collapsed' to the true, more compact structure, while the algorithm typically remains in the stretched 3–D interpretation indefinitely. Thus our experiments lend support to the notion that previous perceptions of 3–D structure constrain future models, but suggest that some modification is required to the incremental rigidity scheme to account fully for human behavior. Note that our perceptual experience is consistent with the experimental observations by Adelson (1985) mentioned earlier, which suggest that even for rigid objects, there is a tendency to perceive objects that are compact in the image as being compact in 3–D space, which can lead to nonrigid perceptions.

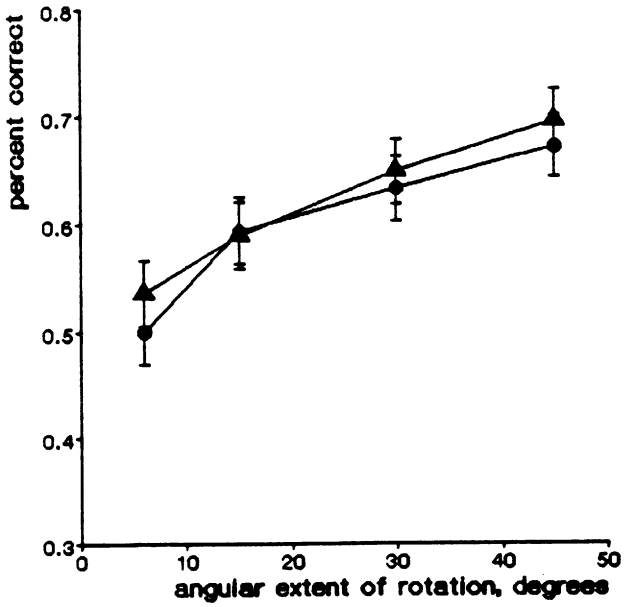

Figure 7. Average data for the experiment in which the test configuration is initially rotated around the line of sight, for subjects ECH and NMG. Circles correspond to the control condition and triangles to the case where the configuration was first rotated around the line of sight. Vertical bars indicate standard errors. There are no significant drops in performance, in contrast to the results obtained for the initial stretched configuration shown in Figure 5.

COMPUTER SIMULATIONS

Through computer simulations, we examined the quantitative behavior of three different formulations of the incremental rigidity scheme at a task similar to that used in the psychophysical experiments. We refer to the first two formulations as Ullman's *discrete model* and *flexible model*, and to the third formulation as the *continuous model*. We first describe the three basic algorithms and then present the results of computer simulations.

Ullman's Discrete Model

Ullman's discrete formulation of the incremental rigidity scheme assumes the visual input to consist of a sequence of frames, each containing a number of discrete points that may correspond to identifiable features in the changing image. The scheme maintains and updates an internal model $M(t)$ of the viewed objects, which consists of a set of 3–D coordinates: $M(t) = (x_i(t), y_i(t), z_i(t))$. All of the formulations used in the simulations here assume orthographic projection onto the $X - Y$ image plane, so that $(x_i(t), y_i(t))$ are the image coordinates of the i–th point, and $z_i(t)$ is the current estimate of the depth at the i–th point (see Grzywacz & Hildreth (1987) for formulations that use perspective projection). When no other 3–D cues are present, the initial

model $M(t)$ at $t = 0$ is taken to be flat; that is, $z_i(0) = 0$ (or some other constant value) for $i = 1, \ldots, n$, where n is the number of points in motion.

Given a current model $M(t)$ at time t and the image of the moving points in a new frame at a later time t', the problem is to compute a new model $M(t')$ such that the transformation from $M(t)$ to $M(t')$ is as rigid as possible. Since $x_i(t')$ and $y_i(t')$ are known, this requires the computation of the unknown depth values $z_i(t')$. (It is assumed that the correspondence between points in the two successive frames is known.) The new depth values are computed as follows. Let $l_{ij}(t)$ denote the distance between points i and j at time t. To make the transformation as rigid as possible, the values $z_i(t')$ for the new model are chosen so as to make $l_{ij}(t)$ and $l_{ij}(t')$ as similar as possible. For this purpose, Ullman defined a measure of the difference between $l_{ij}(t)$ and $l_{ij}(t')$ as:

$$d(l_{ij}(t), l_{ij}(t')) = \frac{(l_{ij}(t) - l_{ij}(t'))^2}{l_{ij}^3(t)}, \tag{1}$$

and formulated the recovery of structure as the computation of $z_i(t')$ that minimize the following overall deviation from rigidity:

$$D_d(t, t') = \sum_{i,j} d(l_{ij}(t), l_{ij}(t')). \tag{2}$$

After the values $z_i(t')$ have been determined using this minimization process, the new model $M(t') = (x_i(t'), y_i(t'), z_i(t'))$ becomes the current model. A new frame is then registered and the process repeats itself. In this way, the scheme maintains rigidity by keeping the total distances between points in the model as constant as possible. The motivation for the cubic factor in the denominator of Equation (1) is that the nearest neighbors to a given point are more likely to belong to the same object than distant neighbors, so that a point is more likely to move rigidly with its nearest neighbors. The $l_{ij}^3(t)$ factor diminishes the influence of distant points in the recovery of structure.

It should be noted that in the case of orthographic projection, only relative depth values, $z_i(t) - z_j(t)$, can be recovered, rather than absolute depth values, because under this form of projection, the image of a given object does not change with its absolute depth. In addition, 3–D structure is determined only up to a reflection about the image plane, since the orthographic projection of a rotating object, and its mirror image rotating in the opposite direction, coincide. Further analysis and variations of this discrete model can be found in Grzywacz and Hildreth (1987).

Ullman's Flexible Model

The flexible model is a modification of the discrete model that allows the internal model at two consecutive instants to be corrected simultaneously. The scheme searches for a modified, corrected model $M'(t)$ such that the transition from $M(t)$ to $M'(t)$ is

small, and the transition from $M'(t)$ to $M(t')$ (the model at time t') is as rigid as possible. The flexible model minimizes the sum:

$$D_f(t,t') = \sum_{i,j} \left[d(l_{ij}(t), l'_{ij}(t)) + d(l'_{ij}(t), l_{ij}(t')) \right] \tag{3}$$

where $l'_{ij}(t)$ refers to the distances between pairs of features in the model $M'(t)$.

The Continuous Model

It is also possible to develop a continuous formulation of the incremental rigidity scheme, which uses velocity information at discrete feature points in a continuously changing image as input to the recovery of 3–D structure (Grzywacz & Hildreth, 1987). We assume again that there always exists an internal model $M(t) = (x_i(t), y_i(t), z_i(t))$, and that the image velocities $\dot{x}_i(t)$ and $\dot{y}_i(t)$ are known. The problem is then formulated as the computation of the z components of velocity, $\dot{z}_i(t)$, that minimize the total continuous change in the distances between the points. The measure of overall deviation from rigidity is given by:

$$D_c(t) = \sum_{i,j} (\dot{l}_{ij}(t))^2 \tag{4}$$

where $\dot{l}_{ij}(t)$ denotes the time derivative of the distances $l_{ij}(t)$, which is dependent on the velocities $(\dot{x}_i(t), \dot{y}_i(t), \dot{z}_i(t))$. The additional factor of l_{ij}^3 that appears in the denomonator of Equation (1) could also be used in the measure shown here. In other respects, the continuous model is similar to Ullman's discrete model. A model of the structure of the moving points is built up by continually taking into account new velocity information over an extended time period. Again, because orthographic projection is used, only relative velocities, $\dot{z}_i(t) - \dot{z}_j(t)$, can be recovered. Further details of the continuous model can be found in Grzywacz and Hildreth (1987).

Simulation Results

This section describes the performance of the three different formulations of the incremental rigidity scheme on a task similar to that used in the psychophysical experiments and compares these simulation results with human performance. Our main conclusion is that the qualitative behaviors of the original discrete model and the continuous model differ significantly from human performance, but the behavior of the flexible model is qualitatively similar to the psychophysical data, at least for the initial rise in performance level for smaller angular rotations.

Configurations of three points were chosen in a way that was similar to the visual stimuli used in the psychophysical experiments, with the positions of the points distributed evenly in depth in the final frame. For each of the three models, the configurations were rotated around the vertical axis for different total angular extents

and images of the points were computed at discrete positions in the trajectory. The simulations of the discrete and flexible models used angular rotations between frames of 10° and a new 3–D model was computed after every 10° of rotation. It has been shown that the performance of the incremental rigidity scheme degrades as the angle of rotation between frames becomes small (Ullman, 1984; Grzywacz & Hildreth, 1987; Landy, 1987). A rotation of 10° was chosen so that we could obtain results for a number of different angular extents in the range from 10° to 180°, without causing significant degradation in the performance of the algorithm. Simulations of the continuous model used angular rotations between frames of 0.1°, so that a new 3–D model was computed every 0.1°. In some simulations, Gaussian distributed noise was added to the positions of the points in each frame. At the end of a given sequence of images, when the frame that contains the points that are equally spaced in depth was reached, it was determined whether the point that occurred in the middle (in depth) in the computed 3–D model was the correct middle point.

In the first set of simulations, the displacements in depth was $\gamma = 40$. For the simulations of the discrete and continuous models, no noise was added to the positions of the moving points. For the simulations of the flexible model, a small amount of Gaussian noise was added, corresponding to $\sigma = 2.0$. Figure 8 shows the percentage of correct responses obtained in the computer simulations, for each of the three models. The results for the discrete and continuous formulations are shown with circles and triangles, respectively, and the results of the flexible model are shown as squares. Each datapoint represents the results of 256 trials. A chance level of performance (33%) was assumed for a rotation of 0°, because the algorithm begins with a flat configuration in which the Z coordinates of the three points are the same. Also shown in the figure is a plot of the psychophysical data from Experiment 2 (crosses)) obtained for $\gamma = 40$, averaged over the three subjects (from Figure 3c).

Some of the degradation in performance seen in this experiment can be attributed to the size of the temporal integration period that is used to

Consider first the behavior of Ullman's discrete model. In the absence of error in the positions of the points, the discrete model eventually converges to a perfect 3–D model and does not reach a plateau at the lower levels of performance seen in the psychophysical experiments. In addition, this model exhibits a slower rise in performance for smaller angular rotations, in comparison with the human data. Thus the discrete model appears not to perform as well as human subjects for small angular rotations, but eventually reaches a significantly better level of performance for larger extents of rotation. We found that a higher level of performance is reached, even for large amounts of added noise in the visual input. The added noise also degrades performance for smaller extents of rotation.

It is interesting to note that early models proposed for recovering structure from motion were not considered viable models for the human recovery of structure, due to

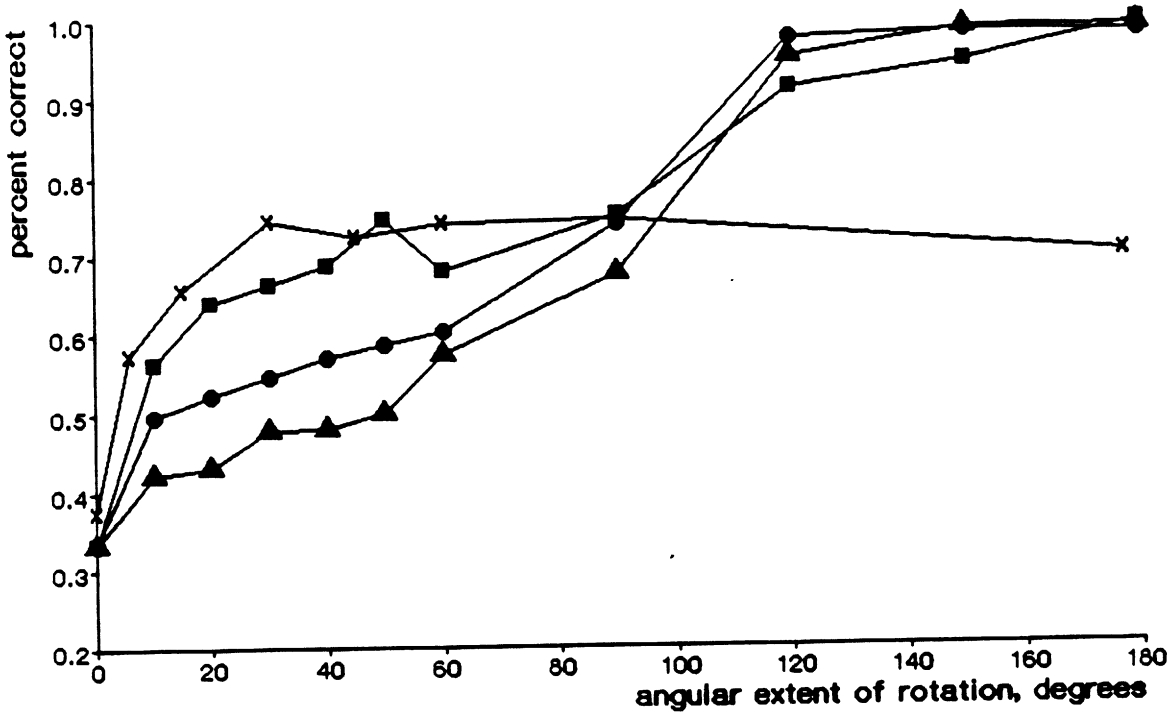

Figure 8. Results of computer simulations. Results of the three models of the incremental rigidity scheme applied to visual stimuli similar to those used in Experiment 2. The graphs correspond to the results of the discrete (circles) and continuous (triangles) formulations, the flexible model (squares) and psychophysical data (crosses). The percentage of correct responses is plotted as a function of the angular extent of rotation. Human performance exhibits a more rapid early buildup in accuracy than the models, but the models continue to improve in performance after the psychophysical data reaches a plateau. The flexible model yields the best fit to the experimental data.

their extreme sensitivity to noise in the visual input. The use of a more flexible rigidity constraint, together with the notion of building up a structure incrementally over time, as proposed by Ullman (1984), has led to an algorithm that can perform better than the human visual system in some circumstances. In order for the discrete formulation of the incremental rigidity scheme to remain viable as a model of human performance, it needs to be modified in a way that yields both better short-term performance and worse performance over extended times. It is possible that some of the differences in observed performance are the consequence of properties of the motion measurement mechanisms preceding the recovery of 3-D structure, which determine the precision of the input position or velocity measurements. Also, these differences may arise from

properties of the way in which the internal model is accessed and the deterioration of its memory over time.

Consider now the continuous model. The study by Grzywacz and Hildreth (1987) showed that the continuous model can provide a good estimate of structure over a short period of time, but then oscillates between good and poor models of structure over an extended time period. It does not yield as stable a long–term recovery of structure as that provided by Ullman's discrete model. From the results shown in Figure 8, it can be seen that in the absence of image noise, the continuous model also reaches an almost perfect level of performance that is significantly higher than the performance levels reached by human subjects. We expect that for larger angular extents of rotation, the performance of the continuous model will drop, because of its oscillatory behavior. The time course of the early buildup of structure is similar to that observed with the discrete model. Simulations showed the continuous model also to be quite robust against noise, but is not as robust as the discrete model. From these simulation results, we conclude that the quantitative behavior of the continuous model also does not appear to agree well with that of human subjects. Similar to the discrete model, it exhibits a slower rise for small angular rotations and eventually reaches a higher level of performance.

The final model that we consider is Ullman's flexible model. In Figure 8, the results for this model are closer to the psychophysical data for angular extents of rotation up to 90°, although human performance is still somewhat better. The flexible model generally builds up a 3–D structure more quickly than the discrete and continuous models. Like the other models, however, it eventually reaches an essentially perfect level of performance for long extents of motion.

We also examined the behavior of the flexible model when a large amount of Gaussian noise is added to the positions of the points in each image frame. Our motivation here is simply to show that similar to human performance, there is a graceful degradation in the behavior of the incremental rigidity scheme with increased noise. Figure 9a shows the performance of the flexible model when Gaussian noise, for which $\sigma = 8.0$, is added to the image frames (squares; shown in comparison to the case in which only a small level of noise is added (circles)). For angular extents less than 90°, there is a drop in performance with the larger amount of added noise, but the algorithm eventually performs at an almost perfect level in both cases. It is difficult to compare, quantitatively, the effects of noise here with the effects observed in the psychophysical experiments, because the nature of the noise in the psychophysical experiments and simulations is different. The simulations use discrete frames at every 10° of rotation, while in the experiments, noise is added to frames that are generated for every 1.5° of rotation. The temporal smoothing that takes place in the early stages of human vision will also tend to smooth out some of the added noise.

Figure 9b shows the results of computer simulations with the flexible model, for

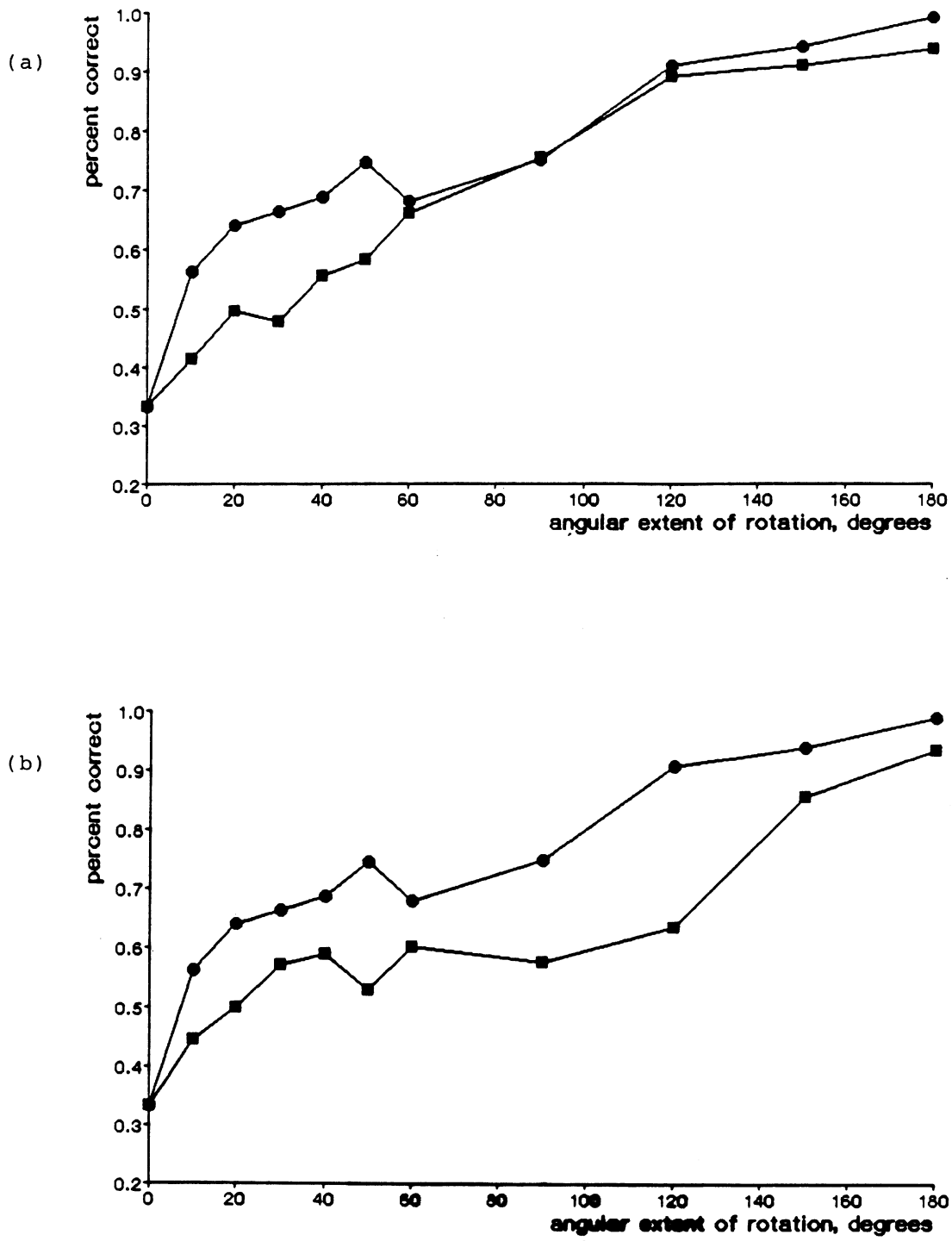

Figure 9. Simulations with the flexible model. (a) The results of the flexible model applied to the visual stimuli used in Experiment 2, with added Gaussian noise in the positions of points in the image frames ($\sigma = 2.0$ (circles) and $\sigma = 8.0$ (squares)). Performance degrades gracefully with additional noise. (b) The results of the flexible model for different displacements in depth, $\gamma = 40$ (circles) and 20 (squares). Performance is worse for the smaller displacement.

two different displacements in depth, $\gamma = 40$ and 20 (shown with circles and squares, respectively). Similar to human behavior, there is an overall drop in performance for the smaller separation in depth.

Overall, we observe qualitative similarities in performance between the flexible model and human subjects for smaller angular rotations (up to 90° or so), but all of the models outperform human subjects for longer rotations. This may be due in part to the difficulty that human observers experienced in maintaining the perception of a rigid configuration of points for long viewing times. It is also possible that human observers are easily confused when the ordering of the points in depth changes during the long angular rotations. If the points were presented in a way that strengthened their apparent rigidity (for example, they were connected with solid lines), there might be an improvement in human performance for longer viewing times.

The key difference between the flexible model and the other two models may be the nature of the updating strategy used at each moment. Allowing the current representation of 3–D structure to change leads to a more rapid early buildup of structure. Note that this updating strategy could be incorporated into either a position–based or velocity–based algorithm.

GENERAL DISCUSSION

This paper presented a series of experiments that assess the accuracy of perceived structure, its sensitivity to noise in the visual image, and the nature of its buildup over time. Our main conclusions are the following. First, the human visual system can derive quite an accurate model of the relative depths of isolated moving points, even in the presence of noise in their image positions. Second, the accuracy of the 3–D model improves with time, eventually reaching a plateau, beyond which there is no further improvement. Third, there is some evidence that the 3–D structure currently perceived depends on previous 3–D models.

The issues of the time course of the buildup of accuracy of perceived 3–D structure and the possible dependence of the currently perceived structure on past 3–D models were specifically motivated by Ullman's incremental rigidity scheme. It is expected that there will be some temporal buildup in accuracy, due to the extended temporal window over which image motion is first measured. The extent of this temporal window, however, is typically on the order of 80−100 msec (for example, McKee & Welch, 1985). Ullman's model proposes that the recovery of structure itself takes place incrementally over a longer time frame. This possibility is supported by our experiments. Subjects showed a buildup in accuracy of perceived structure over a second or so, with some variation between subjects. In quantitative terms, the early time course is similar to that expected by Ullman's flexible model. We also found evidence in our experiments suggesting that the currently perceived structure does depend on past models, although the temporal extent of this effect also may be limited to a second or so.

A limitation of many computational models has been an extreme sensitivity to noise in the visual input. Some models have attempted to overcome this sensitivity by integrating motion measurements at a single moment, but over large spatial areas (for example, Bruss & Horn, 1983; Lawton, 1983; Adiv, 1985; Negahdaripour & Horn, 1985; Ullman, 1984; Waxman & Wohn, 1988), while others overcome this problem through integration of motion measurements over time (Ullman, 1984; Bolles & Baker, 1985; Bharwani et al., 1986; Shariat & Price, 1986; Landy, 1987; Bhanu & Burger, 1988). The results of our experiment with added noise in the visual stimulus suggests that the human system can derive a rough estimate of structure in the presence of large amounts of noise, even when viewing only three points in motion. We may integrate motion information over large spatial regions for some tasks, such as the recovery of observer motion, but an extensive spatial integration by itself cannot account for our experimental observations. This suggests that the integration of motion measurements over time, which may be coupled with viewing the motion over larger spatial extents, may be a more important factor in reducing sensitivity to noise. These noise experiments also suggest that the human visual system may not rely on precise measurements of the velocities and accelerations of image features, but rather may require only rough estimates of the motion of image features, perhaps over an extended time period. This observation is consistent with recent studies of L. Vaina (personal communication) regarding patients with visual deficits, which indicate that patients that lose the ability to make precise velocity discriminations may still be able to recover 3–D structure from motion.

Our experiments couple the total viewing time with the total spatial extent of viewed motion. Other perceptual studies indicate that for extended viewing periods, during which 3–D objects formed from random dots are allowed to oscillate back and forth, the accuracy in perceived 3–D structure increases with the angular extent of rotation (for example, Braunstein et al., 1987; Todd et al., 1988; Loomis & Eby, 1988, 1989). Thus the angular extent of viewed motion is by itself a critical factor in determining the accuracy of perceived structure. Simulations with Ullman's incremental rigidity scheme show that if the spatial extent of rotation of an object is kept constant, but its temporal extent is varied by oscillating the object back and forth through multiple cycles, then the computed 3–D structure will continue to improve over time (Hildreth and Grzywacz, unpublished observations). We conducted a pilot experiment in which our configurations of three points were oscillated back and forth, and did not find any significant improvement in performance over longer times. Thus it remains unclear whether temporal extent of motion is, by itself, a factor in determining the accuracy of computed structure.

While there exists some qualitative similarity between Ullman's incremental rigidity scheme and the human recovery of 3–D structure from motion, there are also clear differences revealed in our experiments. In particular, the somewhat faster early in-

crease in human performance, and the flattening off of performance at a level that is significantly less than perfect pose a challenge that opens the way for further development of the model.

Acknowledgements: We thank Tomaso Poggio, Shimon Ullman and Whitman Richards for valuable comments on this paper, and Mike Landy for useful discussions.

References

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

Adiv, G. (1985). Determining three–dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Machine Intell.*, **PAMI–7**, 384–401.

Andersen, R. A., & Seigel, R. M. (1988). Motion processing in primate cortex. Manuscript submitted for publication.

Barron, J. (1984). A survey of approaches for determining optic flow, environmental layout and egomotion. *Univ. Toronto Tech. Rep. Res. Biol. Comp. Vision*, **RBCV–TR–84–5**.

Bennett, B. M., & Hoffman, D. D. (1985). The computation of structure from fixed axis motion: nonrigid structures. *Biol. Cybern.*, **51**, 293–300.

Bhanu, B., & Burger, W. (1988). Qualitative motion detection and tracking of targets from a mobile platform, Proc. Image Understanding Workshop, Boston, MA, 289–318, San Mateo, CA: Morgan Kaufmann.

Bharwani, S., Riseman, E., & Hanson, A. (1986). Refinement of environmental depth maps over multiple frames, Proc. IEEE Workshop on Motion: Representation and Analysis, Charleston, SC, 73–80.

Bolles, R. C., & Baker, H. H. (1985). Epipolar–plane image analysis: A technique for analyzing motion sequences, Proc. Third IEEE Workshop on Computer Vision: Representation and Control, Bellaire, MI, 168–178.

Borjesson, E., & von Hofsten, C. (1973). Visual perception of motion in depth: application of a vector model to three–dot motion patterns. *Percept. Psychophys.*, **13**, 169–179.

Braunstein, M. L. (1976). *Depth Perception Through Motion.* New York: Academic Press.

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

Braunstein, M. L., & Andersen, G. J. (1984b). Shape and depth perception from parallel projections of three–dimensional motion. *J. Exp. Psych.: Human Percept. Perf.*, **10**, 749–760.

Braunstein, M. L., Hoffman, D. D., Shapiro, L. R., Andersen, G. J., & Bennett, B. M. (1987). Minimum points and views for the recovery of three–dimensional structure. *J. Exp. Psych.: Human Percept. Perf.*, **13**, 335–343.

Bruss, A., & Horn, B. K. P. (1983). Passive navigation. *Comput. Vision Graph. Image Proc.*, **21**, 3–20.

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

Cutting, J. E. (1982). Blowing in the wind: Perceiving structure in trees and bushes. *Cognition*, **12**, 25–44.

Doner, J., Lappin, J. S., & Perfetto, G. (1984). Detection of three–dimensional structure in moving optical patterns. *J. Exp. Psychol.: Human Percept. Perf.*, **10**, 1–11.

Dosher, B. A., Landy, M. S. & Sperling, G. (1989). Ratings of kinetic depth in multi–dot displays. *J. Exp. Psychol.: Human Percept. Perf.*, in press.

Gibson, J. J., & Gibson, E. J. (1957). Continuous perceptive transformations and the perception of rigid motion. *J. Exp. Psychol.*, **54**, 129–138.

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

Grzywacz, N. M., & Hildreth, E. C. (1985). The incremental rigidity scheme for recovering structure from motion: position vs. velocity based formulations. *MIT Artificial Intelligence Laboratory Memo*, **845**.

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

Hildreth, E. C., & Koch, C. (1987). The analysis of visual motion: From computational theory to neuronal mechanisms. *Ann. Rev. Neurosci.*, **10**, 477–533.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *J. Opt. Soc. Am.*, **72**, 888–892.

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

Husain, M., Treue, S., & Andersen, R. (1989). Manuscript submitted for publication.

Jansson, G., & Johansson, G. (1973). Visual perception of bending motion. *Perception*, **2**, 321–326.

Jasinschi, R. & Yuille, A. (1989). Nonrigid motion and Regge calculus. *J. opt. Soc. Amer. A*, **6**, 1088–1095.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.*, **14**, 201–211.

Johansson, G. (1977). Spatial constancy and motion in visual perception. In *Stability and Constancy in Visual Perception*, ed. W. Epstein, New York: Wiley.

Johansson, G. (1978). Visual event perception. In *Handbook of Sensory Physiology*, eds. R. Held, H. W. Leibowitz, H.–L. Teuber, Berlin: Springer–Verlag.

Kanatani, K. (1985). Structure from motion without correspondence: general principle. *Proc. Image Understanding Workshop*, Miami, FL, pp. 107–116.

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

Landy, M. S. (1987). A parallel model of the kinetic depth effect using local computations. *J. Opt. Soc. Amer. A*, **4**, 864–877.

Lappin, J. S., Donner, J. F. & Kottas, B. L. (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.

Lawton, D. T. (1983). Processing translational motion sequences. *Comput. Vision Graph. Image Proc.*, **22**, 116–144.

Longuet–Higgins, H. C., & Prazdny, K. (1980). The interpretation of moving retinal images. *Proc. R. Soc. London Ser. B*, **208**, 385–397.

Loomis, J. M., & Eby, D. M. (1988). Perceiving structure from motion: Failure of shape constancy. *Proc. Second Int. Conf. Comp. Vision*, Tampa, Florida, December, 383–391.

Loomis, J. M., & Eby, D. M. (1989). Relative motion parallax and the perception of structure from motion. *Proc. IEEE Workshop on Visual Motion*, Irvine, CA, 204–211.

McKee, S. P., & Welch, L. (1985). Sequential recruitment in the discrimination of velocity. *J. Opt. Soc. Am. A*, **2**, 243–251.

Mitiche, A. (1986). On kineopsis and computation of structure and motion. *IEEE Trans. Pattern Anal. Machine Intell.*, **PAMI–8**, 109–112.

Negahdaripour, S., & Horn, B. K. P. (1985). Direct passive navigation. *MIT Artif. Intell. Memo* **821**.

Petersik, J. T. (1979). Three–dimensional object constancy: coherence of a simulated rotating sphere in noise. *Percept. Psychophys.*, **25**, 328–335.

Petersik, J. T. (1987). Recovery of structure from motion: Implications for a performance theory based on the structure–from–motion theorem. *Percept. Psychophys.*, **42**, 355–364.

Prazdny, K. (1983). On the information in optical flows. *Comput. Vision Graph. Image Proc.*, **22**, 239–259.

Rashid, R. (1980). LIGHTS: A system for the interpretation of moving light displays. PhD dissertation, University of Rochester.

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

Schwartz, B. J., & Sperling, G. (1983). Nonrigid 3–D percepts from 2–D representations of rigid objects. *Invest. Ophthal. Visual Sci.*, **24**, (3, Supplement), 239.

Shariat, H., & Price, K. E. (1986). How to use more than two frames to estimate motion, Proc. IEEE Workshop on Motion: Representation and Analysis, Charleston, SC, May, 119–124.

Siegel, R. M., & Andersen, R. A. (1988). Perception of three–dimensional structure from two–dimensional visual motion in monkey and man. Manuscript submitted for publication.

Sperling, G., Landy, M. S., Dosher, B. A., & Perkins, M. E. (1989). The kinetic depth effect and identification of shape. *J. Exp. Psych.: Human Perc. Perf.*, in press.

Subbarao, M. (1986). Interpretation of image motion fields: a spatio–temporal approach, Proc. IEEE Workshop on Motion: Representation and Analysis, Charleston, SC, May, 157–165.

Todd, J. T. (1982). Visual information about rigid and nonrigid motion: A geometric analysis. *J. Exp. Psychol.*, **8**, 238–252.

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

Todd, J. T. (1985). The perception of structure from motion: Is projective correspondence of moving elements a necessary condition? *J. Exp. Psych.: Human Percept. Perf.*, **11**, 689–710.

Todd, J. T., Akerstrom, R. A., Reichel, F. D., & Hayes, W. (1988). Apparent rotation in three–dimensional space: Effects of temporal, spatial, and structural factors. *Percept. Psychophys.*, **43**, 179–188.

Tsai, R. Y., & Huang, T. S. (1981). Uniqueness and estimation of three–dimensional motion parameters of rigid objects with curved surfaces. *Univ. Illinois Urbana–Champaign, Coordinated Science Laboratory Report* R–921.

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

Ullman, S. (1983). Computational studies in the interpretation of structure and motion: summary and extension. In *Human and Machine Vision*, ed. J. Beck, B. Hope, A. Rosenfeld. New York: Academic Press.

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

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

Wallach, H., Weisz, A., & Adams, P. A. (1956). Circles and derived figures in rotation. *Am. J. Psych.*, **69**, 48–59.

Waxman, A. M., & Ullman, S. (1985). Surface structure and three–dimensional motion from image flow kinematics. *J. Robotics Res.*, **4**, 72–94.

Waxman, A. M., & Wohn, K. (1988). Image flow theory: A framework for 3–D inference from time–varying imagery. In: *Advances in Computer Vision*, vol. 1, ed. C. Brown, Hillsdale, NJ: Lawrence Erlbaum, 165–224.

Weng, J., Huang, T., & Ahuja, N. (1989). Motion and structure from two perspective views: Algorithms, error analysis, and error estimation. *IEEE Trans. Patt. Anal. Machine Intell.*, **11**, 451–476.

White, B. W., & Mueser, G. E. (1960). Accuracy in reconstructing the arrangement of elements generating kinetic depth displays. *J. Exp. Psychol.*, **60**, 1–11.

Yasumoto, Y., & Medioni, G. (1985). Experiments in estimation of 3–D motion parameters from a sequence of image frames. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, New York: IEEE, 89–94.

CS-TR Scanning Project
Document Control Form

Date : 1/12/95

Report # AIM-1141

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 36 (42 - IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☒ DOD Form (2 PGS) ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages (by page number):_____

Photographs/Tonal Material (by page number):_____

Other (note description/page number):

Description :	Page Number:
(A) PAGES 35 AND 36 ARE WRINKLED	
(B) IMAGE MAP (1-36) PAGES #'S0 1-36 INCLUDING TITLE PAGE	
(37) SCANCONT	
(38-40) TRGT'S	
(41-42) DOD'S	

Scanning Agent Signoff:

Date Received: 1/12/95 Date Scanned: 1/13/95 Date Returned: 1/19/95

Scanning Agent Signature: Michael W. Cook

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**

Scanned

Date: 1/13/1995

M.I.T. Libraries
Document Services

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD# A214 640

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AIM 1141	2. GOVT ACCESSION NO.: AD-A214640
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): The Perceptual Buildup of 3-D Structure from Motion	5. TYPE OF REPORT & PERIOD COVERED: memorandum
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Ellen C. Hildreth, Norberto M. Grzywacz, Edward H. Adelson, Victor K. Inada	8. CONTRACT OR GRANT NUMBER(s): N00014-85-K-0124
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: August 1989 13. NUMBER OF PAGES: 36
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research, Information Systems, Arlington, VA 22217	15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

motion perception
structure from motion
motion analysis
visual perception
3-D vision

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Abstract: We present a set of psychophysical experiments that measure the accuracy of perceived three-dimensional structure derived from relative motion in the changing two-dimensional image. The experiments are motivated in part by a computational model proposed by Ullman (1984), called the *incremental rigidity scheme*, in which an accurate 3-D structure is built up incrementally, by considering images of moving objects over an extended time period. Our main conclusions are: first, the human

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED (cont. on back)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 cont.

visual system can derive an accurate model of the relative depths of moving points, even in the presence of noise in their image positions; second, the accuracy of the 3–D model improves with time, eventually reaching a plateau; and third, the 3–D structure currently perceived appears to depend on previous 3–D models. Through computer simulations, we relate the results of our psychophysical experiments with the predictions of Ullman's model.