Communicated by Christof Koch

Surface Interpolation in Three-Dimensional Structure-from-Motion Perception

Masud Husain
Stefan Treue
Richard A. Andersen
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Although it is appreciated that humans can use a number of visual cues to perceive the three-dimensional (3-D) shape of an object, for example, luminance, orientation, binocular disparity, and motion, the exact mechanisms employed are not known (De Yoe and Van Essen 1988). An important approach to understanding the computations performed by the visual system is to develop algorithms (Marr 1982) or neural network models (Lehky and Sejnowski 1988; Siegel 1987) that are capable of computing shape from specific cues in the visual image. In this study we investigated the ability of observers to see the 3-D shape of an object using motion cues, so called structure-from-motion (SFM). We measured human performance in a two-alternative forced choice task using novel dynamic random-dot stimuli with limited point lifetimes. We show that the human visual system integrates motion information spatially and temporally (across several point lifetimes) as part of the process for computing SFM. We conclude that SFM algorithms must include surface interpolation to account for human performance. Our experiments also provide evidence that local velocity information, and not position information derived from discrete views of the image (as proposed by some algorithms), is used to solve the SFM problem by the human visual system.

1 Introduction

Recovering the three-dimensional (3-D) structure of a moving object from its two-dimensional (2-D) projection is considered an "ill-posed" problem (Poggio and Koch 1985) since there are an infinite number of interpretations of a given 2-D pattern of motion. Several elegant algorithms have been formulated for computing SFM, each using a number of constraints to restrict the number of possible solutions (Ullman 1979, 1984; Longuet-Higgins and Prazdny 1980; Hoffman 1982; Grzywacz and Hildreth 1987). None of them use surface interpolation. Rather these algorithms compute the relative position of isolated points. Existing schemes therefore

require that the tracked points on an object must be present continuously over the entire duration of the SFM computation. This leads to the powerful prediction that if the visual system is forced to sample a new set of points on the same object, the old set of points should not improve the perception of SFM since a new model of the object would have to be computed with each new set of sample points.

An alternative approach to solving the SFM problem is to compute a 3-D surface representation by interpolating a surface between the sample points (Andersen and Siegel 1988). Such a scheme would sample the movement of as many points as possible across the surfaces of the object, and interpolate locally across these measurements to compute a continuous surface. New sets of points can easily be integrated into the representation and thereby improve its accuracy while the information of disappearing points is preserved in the interpolated surface. (We apply the term "surface interpolation" in a general way since the surface could be generated in physical as well as in velocity space.)

2 Experiments

In our experiments we examined how the human visual system performs the SFM computation when confronted with continuously changing sets of sample points. We used novel "structured" and "unstructured" dynamic random-dot stimuli with limited point lifetimes (Morgan and Ward 1980; Zucker 1984; Siegel and Andersen 1988). The structured stimulus was computed from the parallel projection of points covering the surface of a transparent rotating cylinder (Fig. 1). All subjects, whether naive or experienced, have reported the perception of a revolving hollow cylinder when viewing the structured display. The unstructured stimulus was generated by randomly displacing the velocity vectors present in the structured display within the boundaries of the stimulus, thereby conserving the population of vectors but destroying the spatial relationship between them (see Siegel and Andersen 1988). Each point was displayed for a "lifetime" of only 100 msec (7 frames), after which it was replotted randomly at another position on the surface of the cylinder. In the first frame, points were randomly assigned positions in their life cycle. Thus, between two given frames of the stimulus only about 15% of the points "died" and were randomly replotted ("desynchronized case").

Under these conditions, using a reaction time task, we have found that subjects detect the change from an unstructured to a structured display reliably ($> 80\%$ correct) but take as much as 900 msec to react as shown in figure 2. This observation would suggest that the computation of SFM builds up over time and that new points can be integrated into the representation, which is partially computed by the old points. Unfortunately, it is not possible to determine from these data how much of this reaction time is needed as visual input and how much of it is com-

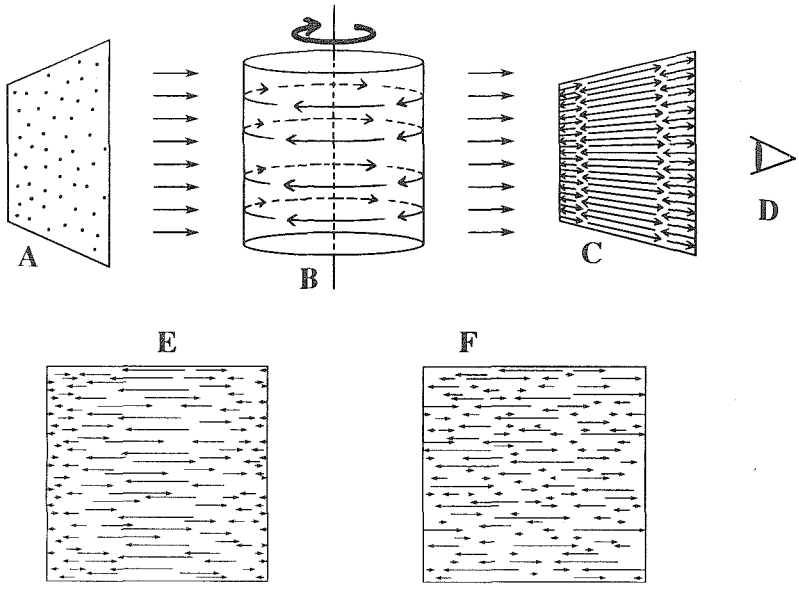

Figure 1: Schematic description of stimulus creation. All movies were created off-line on a PDP 11/73 computer that was also used to run the experiments. For the structured stimulus (E) 126 or 12 points were first randomly plotted on a two-dimensional surface (A). They were then parallel projected onto the surface of a transparent cylinder that was rotated at an angular rotation rate of 35°. sec^{-1} (B). Each point existed for a predetermined point lifetime after which it was randomly repositioned. The moving points were then parallel projected onto the two-dimensional CRT screen (HP 1311B; P31 phosphor) (C) that was viewed by two highly trained observers (D) (ST and MH). The resulting velocity distribution in the structured stimulus is sinusoidal along any horizontal line across the stimulus, with the fastest speeds in the center of the display. The unstructured stimulus (F) was created by randomly shuffling the paths of the points in the structured display. Observers viewed the display binocularly from a distance of 57 cm; the stimuli subtended 6° of visual angle at the eye. The display rate was 70 Hz and mean luminance was 1 cd m^{-2}.

putation time in the brain or motor reaction time. This is an important question since performance should improve when the stimulus is seen for longer than the point lifetime if surface interpolation is used.

2.1 Perceptual Buildup and Surface Interpolation. In order to address this issue we presented equal numbers of structured and unstructured stimuli of 40 to 1700 msec duration in random order and asked subjects to indicate in a two-alternative forced-choice paradigm whether they saw a rotating cylinder or an unstructured noise pattern. Figure 3 (filled symbols) shows that performance peaked only after viewing stimuli longer than 5 times the point lifetime (> 500 msec), being hardly above chance after one point lifetime. Current algorithms (which do not use surface interpolation) would not have predicted improved performance when viewing stimuli of more than one point lifetime.

It could be argued, however, that the visual system selects a number of points from the display and needs to track their relative positions as a group for the duration of their lifetime. Since in our stimulus the points are not synchronized it is very unlikely that all the points in such a group are at the same point in their life cycle, that is, their onsets and offsets do not occur at the same time. So, because groups of dots constantly form

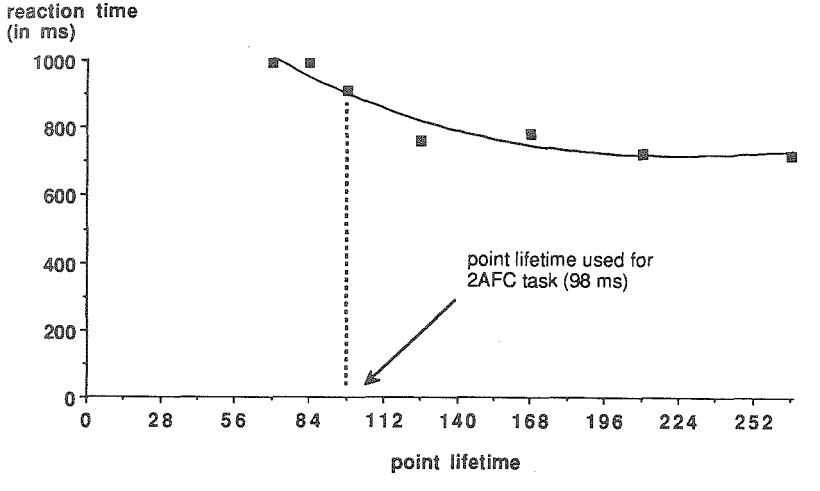

Figure 2: Reaction time for detecting the transition from an "unstructured" to a "structured" cylinder. Observers were shown movies that started with the unstructured version of the cylinder, which after an unpredictable time changed into the structured display. The task was to press a key as soon as the structured cylinder was detected. (For further details see Siegel and Andersen 1988). The arrow and dotted line indicate the point lifetime used for the two-alternative forced-choice experiments described in the text. The regression line is a best-fit third-order polynomial. Each data point represents the mean of about 100 trials.

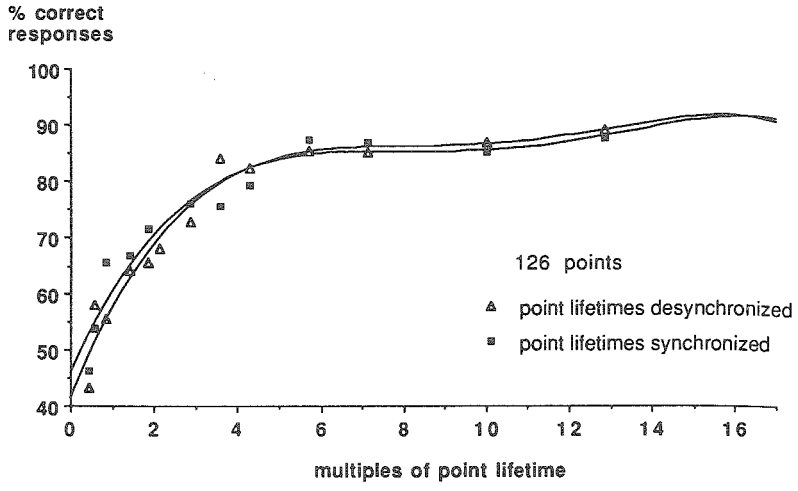

Figure 3: Percentage accuracy in a two-alternative forced-choice paradigm plotted as a function of duration of display in multiples of point lifetime (point lifetime was kept at 100 msec). Observers were shown movies of different duration containing either the cylinder or the unstructured stimulus and were asked to distinguish between them. The dots in the display were either desynchronized (open symbols), or the onsets and offsets of all the dots were synchronized (filled symbols). Note that in both cases peak performance is not reached until over 5 times the lifetime of each point, that is, > 490 msec. The regression lines are best-fit fourth-order polynomials ($r > 0.97$ for both). Each data point represents the mean of 200 to 600 trials.

and dissolve, it might be argued that it simply takes a long time before one finds a group in which all the dots are "in phase." Therefore, we asked our subjects to view displays in which all the points appeared and disappeared together, that is, they were synchronized. Figure 3 (open symbols) shows that performance was indistinguishable from the desynchronized case.

Another important consideration is that a surface interpolation may be used only when the high density of dots in the stimulus already perceptually constitutes a surface. Under different conditions, when the points are not dense enough to constitute an apparent surface by themselves, an alternative algorithm might be used. To investigate if the perceptual buildup we observed occurs only with a high density of dots, we decreased the number of points to less than a tenth of the original 126 points. Figure 4 (open symbols) shows that the time course using 12

points is even longer, with performance peaking only after more than 10 point lifetimes.

To control for the possibility that the buildup in performance is not due to the presentation of new points but to some other effect we performed another experiment. We showed stimuli of the same duration in which the 12 points, after living through their first lifetime, were not randomly replotted but repositioned to the location they originally occupied at the beginning of the movie. They then moved through the same path as before and at the end of their lifetime were again replotted at their original starting position, thus beginning the cycle again. These "oscillating" stimuli therefore contained the same number of points with the same point lifetime as used in the previous experiment but after the passage of the first point lifetime they contained no new information. The results are plotted in figure 4 (filled symbols). It is evident that subjects did not perform above chance under these conditions. Thus, the visual system can improve its performance dramatically when presented with

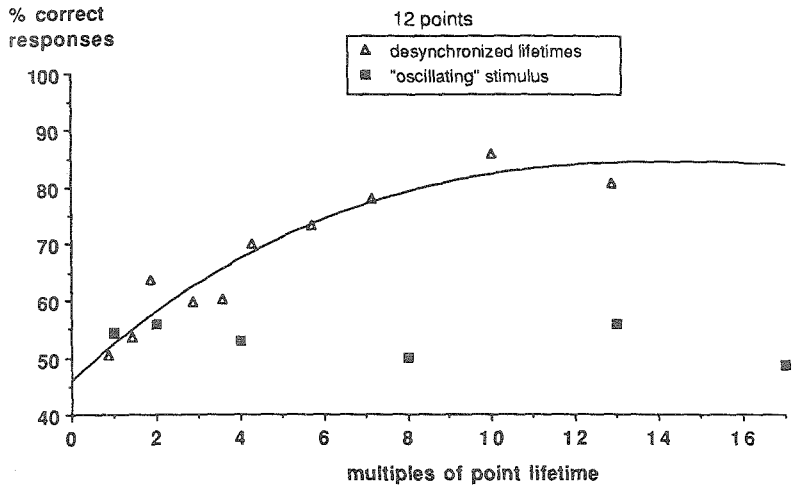

Figure 4: Percentage accuracy plotted as a function of display duration when 12 desynchronized points were used (point lifetime again 100 msec). Open symbols show the results from the experiment comparable to Figure 3. In this case perceptual buildup is more gradual and long-lasting. Peak performance is not reached until a stimulus length of more than 10 point lifetimes, that is, over 1 sec. The regression line is a best-fit third-order polynomial (r = 0.96). Filled symbols show the results from the experiment in which points were replotted to their original position at the end of their point lifetimes (for details see text).

new sets of points and this is not due to a requirement to view stimuli for an extended period of time.

These results suggest that the brain uses surface interpolation in computing the shape of 3-D surfaces from motion. As predicted, the accuracy of the object representation rises to some maximum value with the presentation of new data points, and the performance of the system is not influenced by whether the points are synchronized or not (cf. Fig. 3). Moreover, given less points, it predictably takes longer to compute an accurate surface representation (cf. Fig. 4). As expected, the surface representation integrates information over space, since performance was better with larger numbers of points, and over time, since several point lifetimes were required for the computation.

2.2 Position- versus Velocity-Based Computation. A second issue is whether the visual system measures position or local velocities in computing SFM. Position-based algorithms sample position information derived from a few discrete image views of a moving object and attempt to reach a rigid 3-D interpretation from the 2-D sample frames (Ullman 1984; Grzywacz and Hildreth 1987; Grzywacz *et al.* 1988). Velocity-based algorithms measure the local velocities of points on an image and use the global velocity field to compute 3-D SFM (Longuet-Higgins and Prazdny 1980; Hoffman 1982; Grzywacz and Hildreth 1987).

To date, neither position- nor velocity-based algorithms have used surface interpolation and all velocity-based algorithms have used instantaneous velocity whereas the nervous system requires 50–80 msec to measure velocity accurately (McKee and Welch 1985; Nakayama 1985). A modified position-based scheme could incorporate measurements from new sets of points to improve performance by smoothing over the computed 3-D locations of points to interpolate a surface (E.C. Hildreth and S. Ullman, personal communications).

However, there are several reasons to believe the nervous system uses a velocity-based algorithm with surface interpolation. In our displays the angular extent of the individual movements is quite small, approximately 3.5°, since they are of finite point lifetime. Position-based algorithms require large displacements of 30–50° (Grzywacz and Hildreth 1987). Other experimental support from our laboratory for the velocity-based surface scheme comes from the finding that the minimum point lifetime required for perceiving SFM (Treue *et al.* 1988) corresponds to the minimum viewing time required to measure accurately the velocity of a moving stimulus (McKee and Welch 1985). This correspondence is preserved with changes in stimulus velocity: the point lifetime threshold falls in parallel for both tasks as velocity is increased. This correlation is further strengthened by the fact that subjects can detect motion in our displays with point lifetimes lower than the ones required for comparative performance in detecting SFM, suggesting that the perception of motion per se is not sufficient but that an accurate velocity field has to be measured. Finally,

our laboratory as well as other investigators have shown that lesions of area MT, a region in primate visual cortex that contains neurons tuned to global stimulus direction and velocity (Movshon *et al.* 1985; Allman *et al.* 1985), impair perception of both coherent motion (Newsome and Paré 1988) and SFM (Siegel and Andersen 1986, 1988).

3 General Discussion

There are two possible levels at which a surface interpolation of the velocity field might occur. One is at a 2-D level in which the velocities of points moving on the 2-D retinal image are computed and an interpolation process fills in to form a dense 2-D velocity field from which a 3-D interpretation will be computed by a later process. In the second possibility the 3-D surface is immediately computed from the local 2-D velocities and the interpolation process operates on the 3-D image representation. At present we do not have evidence to distinguish between these two possibilities.

A large number of algorithms for 2-D velocity measurement have been proposed that perform some velocity integration, averaging, or smoothing (Hildreth and Koch 1987; Horn and Schunk 1981; Zucker and Iverson 1986; Yuille and Grzywacz 1988; Bülthoff *et al.* 1989). Some of these algorithms have also been implemented in neural networks (Wang *et al.* 1989). Since all these algorithms integrate motion over local spatial neighborhoods they can account for a number of perceptual phenomena. Unfortunately, they cannot deal with transparent objects such as our rotating cylinder since vectors (with opposing direction) from the front and rear surface would be assigned to one surface, and the averaging of velocities over a patch would yield zero velocity. Evidently, an additional requirement for the successful application of these algorithms to transparent objects is the segregation of surfaces prior to the smoothing operation. For our stimulus, a simple solution is to assign motion in one direction to one surface. To investigate this issue we are presently recording from visual cortex in awake macaque monkeys. Preliminary results indicate that transparent motions in different directions are already separated at the level of VI (Erickson *et al.* 1989).

Acknowledgments

We are grateful to Shabtai Barash, Martyn Bracewell, Roger Erickson, Norberto Gryzwacz, Ellen Hildreth, and Shimon Ullman for their comments on earlier drafts of this manuscript. This work was supported by grants from the NIH, the Sloan Foundation, and the Whitaker Health Sciences Foundation. M.H. is a Harkness Fellow and S.T. is a Fellow of the Evangelisches Studienwerk Villigst, F.R.G. and is supported by the Educational Foundation of America.

References

Allman, J., Miezin, F., and McGuinness, E. 1985. Stimulus specific responses from beyond the classical receptive field. *Ann. Rev. Neurosci.* **8**, 407–430.

Andersen, R.A., and Siegel, R.M. 1989. Local and global order in perceptual maps. In *Signal and Sense*, G.M. Edelman, W.E. Gall, and W.M. Cowan, eds., in press. Wiley, New York.

Bülthoff, H., Little, J., and Poggio, T. 1989. A parallel algorithm for real-time computation of optical flow. *Nature (London)* **337**, 549–553.

De Yoe, E.A., and Van Essen, D.C. 1988. Concurrent processing streams in monkey visual cortex. *Trends Neurosci.* **11**, 219–226.

Erickson, R.G., Snowden, R.J., Andersen, R.A., and Treue, S. (in press). Directional neurons in awake rhesus monkeys: Implications for motion transparency. *Soc. Neurosci. Abst.*

Grzywacz, N.M., and Hildreth, E.C. 1987. Incremental rigidity scheme for recovering structure from motion: Position-based versus velocity-based formulations. *J. Opt. Soc. Am.* **4**, 503–518.

Grzywacz, N.M., Hildreth, E.C., Inada, V.K., and Adelson, E.H. 1988. The temporal integration of 3-D structure from motion: A computational and psychophysical study. In *Organization of Neural Networks*, W. von Seelen, G. Shaw, and U.M. Leinhos, eds., pp. 239–259. VCH, Weinheim.

Hildreth, E.C., and 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.

Horn, B.K.P., and Schunk, B.G. 1981. Determining optical flow. *Artificial Intelligence* **17**, 185–203.

Lehky, S.R., and Sejnowski, T.J. 1988. Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature (London)* **333**, 452–454.

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

Marr, D. 1982. *Vision*. Freeman, San Francisco.

McKee, S.P., and Welch, L. 1985. Sequential recruitment in the discrimination of velocity. *J. Opt. Soc. Am.* **A2**, 243–251.

Morgan, M.J., and Ward, R. 1980. Interocular delay produces depth in subjectivity moving noise patterns. *Q. J. Exp. Psychol.* **32**, 387–395.

Movshon, J.A., Adelson, E.H., Gizzi, M.S., and Newsome, W.T. 1985. The analysis of moving visual patterns. In *Pattern Recognition Mechanisms* (Exp. Br. Res. Suppl. 11), C. Chagass, R. Gattas, and C. Gross, eds., pp. 117–151. Springer-Verlag, Heidelberg.

Nakayama, K. 1985. Biological image motion processing: A review. *Vision Res.* **25**, 625–660.

Newsome, W.T., and Paré, E.B. 1988. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.* **8**, 2201–2211.

Poggio, T., and Koch, C. 1985. Ill-posed problems in early vision: From computational theory to analog networks. *Proc. R. Soc. London Ser. B* **226**, 303–323.

Siegel, R.M. 1987. A parallel distributed processing model for the ability to obtain three-dimensional structure from visual motion in monkey and man. *Soc. Neurosci. Abstr.* **13**, 630.

Siegel, R.M., and Andersen, R.A. 1986. Motion perceptual deficits following ibotenic acid lesions of the middle temporal area in the behaving rhesus monkey. *Soc. Neurosci. Abstr.* **12**, 1183.

Siegel, R.M., and Andersen, R.A. 1988. Perception of three-dimensional structure from motion in monkey and man. *Nature (London)* **331**, 259–261.

Treue, S., Husain, M., and Andersen, R.A. 1988. Human perception of 3-D structure from motion: Spatial and temporal characteristics. *Soc. Neurosci. Abstr.* **14**, 1251.

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

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

Wang, H.T., Mathur, B., and Koch, C. 1989. Computing optical flow in the primate visual system. *Neural Comp.* **1**, 92–103.

Yuille, A.L., and Grzywacz, N.M. 1988. A computational theory for the perception of coherent visual motion. *Nature (London)* **333**, 71–74.

Zucker, S.W. 1984. Type I and Type II processes in early orientation selection. In *Figural Synthesis*, P.C. Dodwell and T. Caelli, eds., 283–300. Lawrence Erlbaum, London.

Zucker, S.W., and Iverson, L. 1986. From orientation selection to optical flow. Memo CIM-86-2, Computer Vision and Robotics Laboratory, McGill Research Center for Intelligent Machines.

Received 17 April 1989; accepted 16 June 1989.