Detection of one versus two objects in structure from motion

Jeffrey C. Liter, Myron L. Braunstein, and Donald D. Hoffman

Department of Cognitive Sciences, University of California, Irvine, Irvine, California 92717

Received August 19, 1993; revised manuscript received July 18, 1994; accepted July 18, 1994

The ability of subjects to detect whether a structure-from-motion display depicts one or two rigid objects was examined in the presence or the absence of noise points. Each object was composed of a set of points chosen randomly within the volume of a sphere. The objects rotated rigidly about different axes passing through the center of the sphere. For displays without noise points, detection increased with larger angles between the rotation axes and with more points in each object. For displays in which noise points were present, detection was above chance but, in general, worse than that for displays without noise points. The implications of these results for image segmentation in complex motion patterns is discussed.

1. INTRODUCTION

To be recognized, an object must first be separated, or segmented, from other objects in the scene. Many sources of information facilitate segmentation: notably luminance, color, and texture variations.[1] Even without these variations motion alone may reveal the correct segmentation.[2] Most theoretical and experimental investigations of motion segmentation have studied differences in direction or speed of image motion.[3–5] Regions are segmented if they differ in either direction or speed of two-dimensional (2-D) motion. We study rigid motion in three dimensions as a basis for segmentation. For an object rotating about an axis not in the image plane the 2-D projections of the features on its surface trace elliptical paths. Thus the direction of motion is not constant even for a single feature. Likewise, the projected speeds of the object's features are not constant. The image speed of each feature varies sinusoidally through the course of a rotation, and its maximum speed depends on its distance from the axis. Segmentation of motion into regions with common 2-D direction or common 2-D speed would be ineffective for these stimuli.

Bennett *et al.*[6] described a simple computation to determine whether features moving in two dimensions are part of the same rigid three-dimensional (3-D) structure. Briefly, the image coordinates of four features in two time frames are substituted into a polynomial. If the polynomial evaluates to zero, then the features are part of the same rigid 3-D structure (up to a measure zero set of false targets). By applying this algorithm to all features of an image (perhaps in parallel), one can decompose the image into rigid objects. Furthermore, Bennett *et al.* showed that the polynomial degrades gracefully if the position of each image feature is perturbed by Gaussian noise. Their analysis demonstrates that segmentation on the basis of rigid motion in three dimensions is possible in principle and is potentially robust.

In the experiments that follow, subjects were shown a collection of moving dots that simulated the motion of either one or two rigidly rotating objects. The objects were filled transparent spheres with the same radius and center (see Fig. 1). We employed transparent overlapping spheres to avoid motion-segmentation cues other than rigidity. There was no lateral separation of the objects, which would have produced nonrigid boundaries between the objects. The objects were transparent, so that there was no dynamic occlusion. Finally, there were no simple differences in the 2-D motion direction or speed of the objects' features. When two objects were simulated, these controls yielded configurations in which the objects interpenetrated in three dimensions. The task of the subjects was to determine whether one or two objects was present. In experiment 1 each dot in the display was rigidly connected to one of the objects. In experiment 2 additional noise dots were present that did not move rigidly with either object.

2. EXPERIMENT 1

A. Method

The subjects were the first author (JCL) plus three students from the University of California, Irvine, who were naïve to the purposes of the experiment. The naïve subjects were paid for their participation. All the subjects had normal or corrected-to-normal vision (20/40, Snellen eye chart).

Two independent variables were examined in this experiment: the number of points in each object (4, 11, or 32) and the 3-D angle between the two rotation axes (0°, 2°, 4°, 6°, 8°, 10°, 12°, or 14°). Both variables were run within subjects. Forty responses were collected for each nonzero axis separation, and an equal number of responses (280 for each number of points) was collected for the zero-separation condition.

The stimuli were white dots moving on a black background. The motion of the dots simulated orthographic projection of either one rotating object (zero axis separation) or two rotating objects (nonzero axis separation) overlapping in three dimensions. To produce the motion, we chose points randomly within the volume of a sphere of radius 512 pixels (Ref. 7) such that no point was within 102 pixels of the center of the sphere. The sphere was centered at the origin of a Cartesian coordinate system in

0740-3232/94/123162-05$06.00

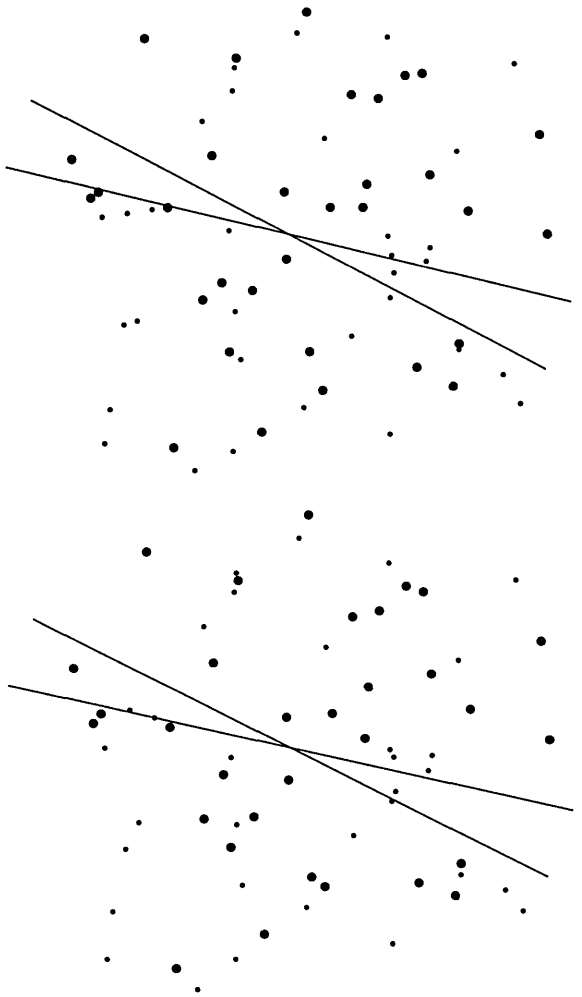

Fig. 1. Stereo illustration of a SFM display in experiment 1. The reader should view the stereogram by turning the page sideways. There are 32 points in each object. We used different-sized dots in this illustration to differentiate the two objects. (Only one dot size was used in the actual displays.) The solid lines, not present in the actual displays, indicate the two rotation axes. The 3-D angle between the rotation axes in this display is 14°.

which Y was vertical, X was horizontal, and Z was coincident with the line of sight. Two rotation axes passing through the origin were selected for each trial. Half of the points were selected randomly to rotate about one axis, and the other half were selected to rotate about the other. The 3-D angle between the two axes, the separation angle, was varied across trials but was kept fixed for a given trial.

The initial direction of the two axes was selected randomly with the constraint that the slant (the angle between the axis and the image plane) of each axis not exceed 15°. A new direction for the axes was selected on each frame transition. The slant of the vector normal to the two axes was not permitted to change by more than 3°, and the tilt was not permitted to change by more than 10° from one frame to the next. These restrictions yielded an average 3-D change of approximately 2.65° per frame in the direction of the axes. The slant of neither axis was permitted to exceed 15° at any time. There were 150 frames in each display. The objects rotated at a rate of 1.5° per frame, and the update and refresh rates were 30 Hz.

The displays were presented on a 21-in. CRT (Xytron model AB21 with P4 phosphor) with 4096 × 4096 resolution. The CRT was controlled by a Vax Station II computer. The subject sat in a completely darkened room 110 cm from the CRT and viewed the displays monocularly through a reduction tube. The tube was fitted with a 0.6 neutral-density filter so that traces on the CRT would not be visible. The display was 1024 pixels in diameter and subtended a visual angle of 5.7°. The subject responded by pressing one of two telegraph keys. Response times were collected by the computer. The timer started when the first dot was presented on the CRT and stopped when either response key was pressed.

Each subject participated in twelve 40-min sessions. Each session consisted of six blocks of trials—two blocks at each numerosity level presented in a random order. Each block contained two repetitions of each of the seven nonzero axis separations plus an equal number of zero-separation conditions. The first block of each session was preceded by six practice trials, and the remaining five blocks were each preceded by four practice trials. The first two experimental sessions were practice, and data from these sessions were not included in the analyses.

Subjects were instructed to determine on each trial whether the displayed motion depicted one or two rigidly rotating objects. The subjects were informed that their response times would be collected but that accuracy should be their primary concern. Each display lasted up to 5 s. If the subject responded in less than 5 s, the display ended and the next trial began. The time between trials was 5 s.

B. Results and Discussion

Each subject's ability to detect the presence of two objects was assessed by the computation of a d' score for each nonzero level of axis separation at each level of point numerosity. One false alarm rate was computed for each level of point numerosity based on the zero axis separation trials at that numerosity level. The d' scores for each subject are presented in Fig. 2. A two-way (7 axis separations by 3 dot numerosities) repeated-measures analysis of variance was conducted on the d' scores. The effect of axis separation was significant [$F(6, 18) = 18.08$, $p < 0.05$, and $\omega^2 = 0.623$] and indicated that detection increased with greater separation. This result obtained for all three levels of point numerocity [$F(6, 18) = 10.03$, 11.94, and 19.26, respectively, for 4, 11, and 32 points per object, all with $p < 0.05$]. The effect of point numerosity was also significant [$F(2, 6) = 10.45$, $p < 0.05$, and $\omega^2 = 0.143$]. A significant interaction between axis separation and point numerosity revealed that detection increased faster with axis separation for objects with more points [$F(12, 36) = 2.89$, $p < 0.05$, and $\omega^2 = 0.024$].

For the most part, subjects responded shortly after the 5-s display ended. For one-object responses (whether correct or incorrect) the response time was constant regardless of axis separation. For two-object responses the response times decreased slightly for greater axis separations.

It is clear that subjects can determine whether a structure-from-motion (SFM) display contains one or two

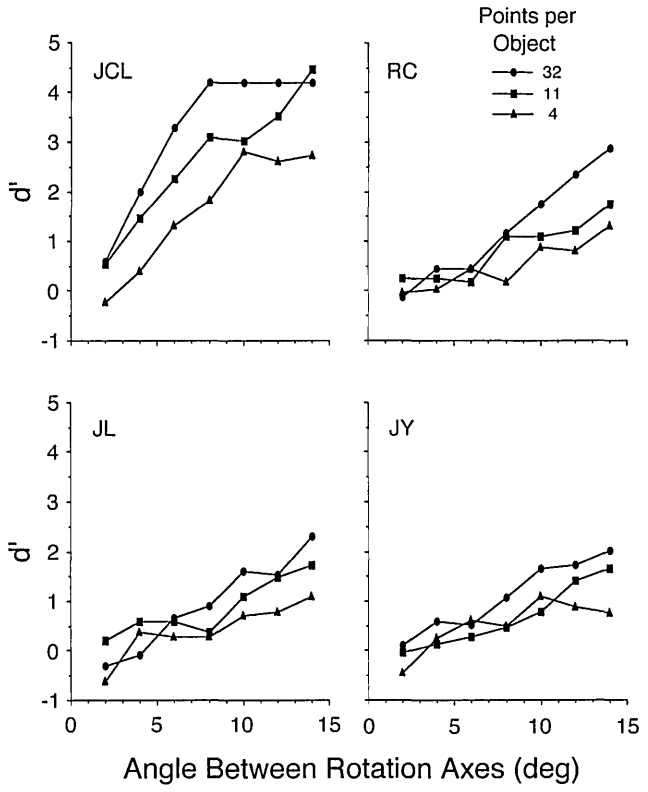

Fig. 2. Detection (d') as a function of rotation axis separation for the three levels of dot numerosity in experiment 1.

rigid objects. It is not clear, however, whether our subjects were actually segmenting the displays into two objects. Since only the two-object displays contained 3-D nonrigidity, it is possible that subjects made their judgments on the basis of nonrigidity. The increase in performance with greater point numerosity is consistent with this hypothesis, since the ability to detect nonrigidity should increase with the number of nonrigid motions present in the display. In addition, none of the subjects reported a percept of two distinct objects in the two-object displays. In contrast, a perception of separate surfaces is usually found in motion transparency experiments.[8] All the subjects did, however, report a vivid 3-D percept on all the trials.

3. EXPERIMENT 2

For segmentation by rigidity to be useful within the framework of object recognition, it must go beyond simple detection of nonrigidity; it must group features into distinct subsets that correspond to distinct objects in the scene. The results of experiment 1 indicate that subjects can use the presence of nonrigidity to determine whether a SFM display contains one or two rigid objects. In experiment 2 we include nonrigid points in all the displays to ensure that the task cannot be performed by simple detection of nonrigidity.

A. Method

The subjects were the first author (JCL) and one paid graduate student (JSK) from the University of California, Irvine, who was aware of the purposes of the experiment. Both subjects had normal or corrected-to-normal vision.

Two independent variables were examined: the number of points in each object (4 or 12) and the 3-D angle between the rotation axes of the two objects (0° or 24°). Both variables were run within subjects.

The stimuli were similar to those used in experiment 1 except that each display contained as many noise points as there were points in each object (i.e., 4 or 12 noise points). Noise points were selected in the same way as were object points, but each noise point rotated about a unique axis. (Note that this type of noise is different from the addition of Gaussian noise to individual point coordinates studied by Bennett *et al.*) The noise axes were selected randomly with the following constraints. When the angle between axes was 0° (i.e., one object was simulated), the 3-D angle between that axis and each noise rotation axis was 24°. Thus, in this case, all the noise axes fell on a single cone with radius 24°. When the angle between the two object axes was 24° (i.e., two objects were simulated), the noise rotation axes were selected from two cones, one surrounding each object axis. Each of these cones, however, had a radius of 12°. This procedure for selecting rotation axes for the noise points yielded similar 3-D nonrigidity[9] for the one-object and two-object displays. In a Monte Carlo simulation the 3-D nonrigidity [mean (standard deviation) of 40 repetitions] for 4-point displays was 26.7 (4.8) and 32.4 (5.4) pixels for one-object and two-object displays, respectively. The 3-D nonrigidity for 12-point displays was 27.4 (2.7) and 31.7 (3.0) for one-object and two-object displays, respectively.

As in experiment 1, the directions of all the axes, including noise axes, were updated on every frame transition, but the 3-D angle between each pair of axes was kept constant. Unlike the procedure in experiment 1, each 5-s display repeated after a 1-s blank interval, so that the subject had more time to view the display. If no response was made before 60 s, the trial ended and was repeated later in the session. The apparatus was the same as that in experiment 1.

Each subject participated in 12 experimental sessions. The sessions were blocked by the number of points in each object. Each session consisted of two blocks of trials, 5 practice trials followed by 20 completely randomized experimental trials (10 one-object trials and 10 two-object trials). Subjects did not receive feedback in the first four sessions but did receive feedback in the remaining eight sessions. The order of sessions was counterbalanced within and between subjects with respect to number of points.

B. Results and Discussion

The detection data for both subjects are presented in Fig. 3. Data from the first four sessions in which no response feedback was provided are presented in the two leftmost columns of each panel. Data from the final eight sessions in which response feedback was provided are presented in the middle (sessions 5–8) and rightmost (sessions 9–12) columns. All the d' scores are significantly greater than zero ($p < 0.05$), except for the 12-point condition with no feedback for subject JSK. It is not possible to determine for that condition whether the performance improvement in later sessions was due to general practice or to the use of feedback. However, there does not appear to have been a general benefit of feedback.

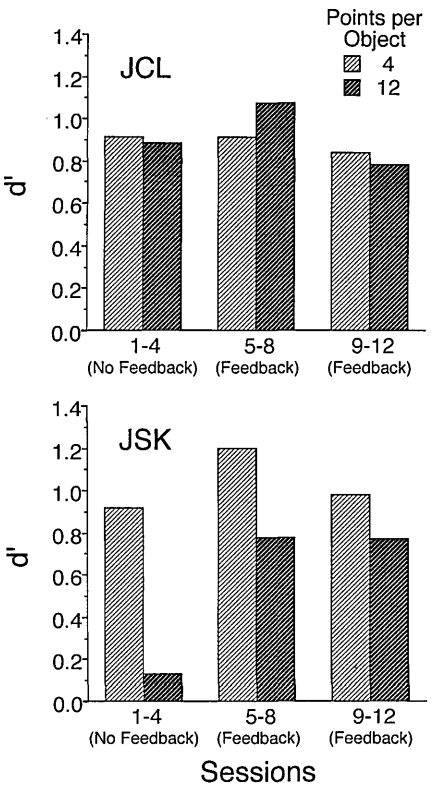

Fig. 3. Detection (d') for objects with 4 points (light bars) and 12 points (dark bars) with an equal number of noise points (experiment 2). Feedback was provided in sessions 5–12 only.

Overall, detection was worse here than in experiment 1, in which no noise points were present in the displays. This is further evidence that, in experiment 1, the presence of nonrigidity in the two-object displays permitted subjects to respond accurately. Detection, however, was above chance in the present experiment. Thus there was information in these displays, beyond the presence of nonrigidity, that subjects used to respond at better-than-chance levels. As in experiment 1, subjects reported that the two-object displays were not perceived as segregated into two objects.

4. GENERAL DISCUSSION

The primary focus of these experiments was to determine whether observers could segment a SFM display on the basis of 3-D rigid motion. The results indicate that subjects can determine whether a display contains one or two objects in the absence of noise and can perform at above-chance levels even in the presence of noise. However, in neither experiment did subjects report a percept of separate objects in two-object displays. As we discussed in Section 1, the stimuli used in the present experiments were specifically designed to eliminate other potential motion-segmentation cues, such as lateral separation of the objects and dynamic occlusion. These controls yielded configurations in which the objects interpenetrated in three dimensions. One explanation of why segmentation was not evident is that the visual system is incapable of seeing two rigid objects move through each other. Instead, it interprets such a motion as a single nonrigid object. This hypothesis is consistent with the subjects' verbal reports in both experiments. All the subjects reported having a 3-D percept, but none perceived the motion as two distinct objects. Another explanation is that motion segmentation is restricted to simple 2-D differences in direction or speed of motion. In research conducted thus far, neither of these explanations can be eliminated.

Subjects' inability to segment these displays was most likely not due to the presence of overlapping velocity fields. Transparent objects in which the front and back surfaces are simultaneously visible are easily perceived as such.[10] Similarly, subjects can easily perceive superimposed transparent surfaces as distinct. Andersen,[8] studying motion parallax stimuli, found that subjects could detect as many as three overlapping planes translating either perpendicular to or along the line of sight. De Bruyn and Orban[11] found that subjects could simultaneously determine the rotation direction (clockwise or counterclockwise) of a disk rotating in the image plane and the direction of motion (toward or away from the observer) of a disk translating along the line of sight. In all these examples, however, the different surfaces occupied different depth planes. Thus no interpenetration occurred. Furthermore, all contained simple 2-D motion-segmentation cues. For a single transparent object the front and back move in different directions if it is rotating about its center and at different speeds if it is translating. For Andersen's motion parallax stimuli the speeds of nearby points in the image were proportional to their simulated distances from the observer. Thus segmentation by image speed was possible. For De Bruyn and Orban's stimuli the direction of motion of the image points with respect to the center of the display was different for the two superimposed surfaces. The features on the translating disk moved toward or away from the center of the display, and the features on the rotating disk moved in circular paths around the center of the display.

Andersen and Wuestefeld[12] examined the detection of smooth surfaces embedded in noise for motion parallax displays. In their experiment 5 they found that detection of a sinusoidal surface was better if the noise points were separated in depth from the surface than if they overlapped the surface. Detection, however, was above chance for many of the conditions in which the noise overlapped the sinusoidal surface. It is important to note that, even in the overlapping noise conditions, the entire configuration (surface plus noise) moved rigidly in three dimensions. Even though the surfaces intersected, they did not move through one another. To examine fully the effect of interpenetration on segmentation, one will need to study nonrigid configurations.

Finally, we note the possibility that other segmentation information might need to be present for rigidity to be used. Rigidity alone may not reveal the proper segmentation but might sort out ambiguities or inconsistencies left by other segmentation information.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation grants DBS-9209773 and DIR-9014278 and U.S. Office of Naval Research contract N00014-88-K-0354. The authors would like to thank J. Turner, K. Jeon,

J. Kim, and A. Saidpour for helpful discussions and J. Szutu for assistance in conducting the experiments. The results of these experiments were presented at the national meeting of the Association for Research in Vision and Ophthalmology, May 1993, Sarasota, Florida.

Correspondence should be addressed to Jeffrey C. Liter, Department of Cognitive Sciences, School of Social Sciences, University of California, Irvine, California 92717-5100. e-mail: jliter@aris.ss.uci.edu.

REFERENCES AND NOTES

1. J. Beck, "Textural segmentation," in *Organization and Representation in Perception,* J. Beck, ed. (Erlbaum, Hillsdale, N.J., 1982), pp. 285–317.
2. The classic example of this phenomenon was provided by H. V. Helmholtz, *Physiological Optics* (MIT Press, Cambridge, Mass., 1910/1962), Vol. III.
3. D. Regan and K. I. Beverley, "Figure-ground segregation by motion contrast and by luminance contrast," J. Opt. Soc. Am. A **1,** 433–442 (1984).
4. C. L. Baker and O. J. Braddick, "Does segregation of differently moving areas depend on relative or absolute displacement?" Vision Res. **22,** 851–856 (1982).
5. A. B. Sekuler, "Motion segregation from speed differences: evidence for nonlinear processing," Vision Res. **30,** 785–795 (1990).
6. B. M. Bennett, D. D. Hoffman, and C. Prakash, "Recognition polynomials," J. Opt. Soc. Am. A **10,** 759–764 (1993).
7. The term pixel does not strictly apply to the calligraphic displays used in these experiments but is used as a convenient shorthand for plotting position, the number of distinct locations at which the center of a dot can be positioned. We use the same pixel units in the z (depth) dimension to describe the geometric model that produces the displays.
8. G. J. Andersen, "Perception of three-dimensional structure from optic flow without locally smooth velocity," J. Exp. Psychol. Human Percept. Perform. **15,** 363–371 (1989).
9. We used the measure of 3-D nonrigidity defined by M. L. Braunstein, D. D. Hoffman, and F. E. Pollick, "Discriminating rigid from nonrigid motion: minimum points and views," Percept. Psychophys. **47,** 205–214 (1990). For each pair of points the variance in their 3-D interpoint distance was computed across all 150 frames of the motion sequence. These values were then averaged to yield a single measure for the display.
10. M. L. Braunstein, "Perceived direction of rotation of simulated three-dimensional patterns," Percept. Psychophys. **21,** 553–557 (1977).
11. B. De Bruyn and G. A. Orban, "Segregation of spatially superimposed optic flow components," J. Exp. Psychol. Human Percept. Perform. **19,** 1014–1027 (1993).
12. G. J. Andersen and A. P. Wuestefeld, "Detection of three-dimensional surfaces from optic flow: the effects of noise," Percept. Psychophys. **54,** 321–333 (1993).