Perception, 1999, volume 28, pages 725 – 737

DOI:10.1068/p2925

Stored representations of three-dimensional objects in the absence of two-dimensional cues

Raymond E Phinney§, Ralph M Siegel¶
Center for Molecular and Behavioral Neuroscience, Rutgers University, 197 University Avenue, Newark, NJ 07102, USA; e-mail: axon@cortex.rutgers.edu
Received 21 July 1998, in revised form 15 December 1998

Abstract. Object recognition was studied in human subjects to determine whether the storage of the visual objects was in a two-dimensional or a three-dimensional representation. Novel motion-based and disparity-based stimuli were generated in which three-dimensional and two-dimensional form cues could be manipulated independently. Subjects were required to generate internal representations from motion stimuli that lacked explicit two-dimensional cues. These stored internal representations were then matched against internal three-dimensional representations constructed from disparity stimuli. These new stimuli were used to confirm prior studies that indicated the primacy of two-dimensional cues for view-based object storage. However, under tightly controlled conditions for which only three-dimensional cues were available, human subjects were also able to match an internal representation derived from motion to that of disparity. This last finding suggests that there is an internal storage of an object's representations in three dimensions, a tenet that has been rejected by view-based theories. Thus, any complete theory of object recognition that is based on primate vision must incorporate three-dimensional stored representations.

1 Introduction

As we go about our daily lives, we receive constantly changing visual information. Not only do we see novel objects and images, but we see familiar objects from new perspectives. The retinal shape of a single object from two different viewpoints can vary radically. So, how do we compare the variable sensory representation of an object to internally stored visual representations?

There are two general hypotheses concerning how three-dimensional objects are compared to an internal representation. The first entails the comparison of a stored three-dimensional representation (Shepard and Metzlar 1971; Marr and Nishihara 1982; Biederman 1987; Ullman 1989) to the incoming sensory representation. The second hypothesis proposes that there is a small set of stored internal two-dimensional 'snapshots' that are distorted to match the incoming sensory representation (Poggio and Edelman 1990; Bülthoff and Edelman 1992; Logothetis and Pauls 1995). The former hypothesis is often called object-centered object recognition and the latter is often called viewer-centered or view-based object recognition (Ullman 1979).

Most object-centered theories postulate that a single three-dimensional object representation is formed and that this single representation holds all of the necessary information to mediate object recognition due to the use of the three-dimensional information to construct a mental 'solid' (Biederman 1987; Marr and Nishihara 1982). The view-based theories postulate that a limited number of two-dimensional views of familiar objects are stored (Logothetis and Pauls 1995). Object recognition is achieved by searching for the closest match between the two-dimensional stored representation and the current sensory representation. This is sometimes modified to also include affine transforms of the stored two-dimensional views to achieve the match (Vetter et al 1995). View-based theories tend to be supported by two types of evidence. First, the addition of depth

§ Current address: Department of Cell Biology, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA
¶ Author to whom all correspondence and requests for reprints should be addressed.

information through binocular disparity and rotation in depth fails to improve object recognition performance over that seen with pure two-dimensional stimuli (Edelman and Bülthoff 1992; Blicher 1995; Bülthoff et al 1995). Second, a single view of an object can, under certain circumstances, support object recognition for up to a 40° rotation of the object (Logothetis and Pauls 1995).

However, since two-dimensional and three-dimensional cues are largely redundant in that they delineate the same object, it should be no surprise that adding three-dimensional cues yields only small performance increases. By analogy, one might make the incorrect assertion that human subjects cannot see the motion of texture-defined objects by noting that experimental judgments of motion direction or speed do not improve greatly when texture is added to luminance elements. The true test of the assertion is to evaluate performance with stimuli defined by motion texture alone.

In an analogous fashion, one must first be able to exercise independent control of two-dimensional and three-dimensional shape cues before declaring the relative contributions of two-dimensional or three-dimensional cues to the process. Prior studies always tested two-dimensional cues alone, or two-dimensional plus three-dimensional cues (Edelman and Bülthoff 1992; Blicher 1995; Bülthoff et al 1995). If purely three-dimensional cues can support object recognition in the complete absence of two-dimensional cues, then object recognition is by definition not a pure two-dimensional process. Issues of how these two attributes interact can then be addressed by studies similar to the three studies listed above. In general, these studies have explored object recognition within a visual submodality. Thus, subjects may be asked to compare or match two objects, both defined by luminance contours.

The current study takes two novel approaches. First, it requires subjects to learn an internal representation of an object in one visual submodality and compare it to a representation obtained from a different submodality. Here motion and disparity cues are used as they can be carefully matched in many stimulus dimensions and they can be manipulated to prevent a subject from using low-level, non-shape cues. The second novel approach lies in the ability to manipulate two-dimensional and three-dimensional object cues directly and independently. This study presents a new method that completely removes all two-dimensional information (eg explicit two-dimensional luminance cues), leaving only three-dimensional cues. To achieve this more rigorous test, novel motion-based and disparity-based stimuli in which three-dimensional and two-dimensional form cues were computer generated from collections of randomly placed flickering dots were used. These two approaches were then used to test whether a purely three-dimensional internal representation can be formed. Human subjects were indeed able to utilize purely three-dimensional cues to identify objects in a forced-choice paradigm, suggesting that a three-dimensional internal representation is both encoded and utilized in primate cortex. These novel stimuli should prove effective in examining the issues of how objects are segmented at the neuronal and network level.

2 Methods

2.1 *Behavioral task*

Subjects were shown the two stimuli separated by a 1 s interstimulus interval and reported whether the shapes of the two objects were the same or different by pressing one of two keys (figure 1a). The first object, or 'standard', was always defined by structure-from-motion (Wallach and O'Connell 1953; Ullman 1979; Siegel and Andersen 1988), and rotated in depth through 360° over 6 s. There were no disparity cues in this display, henceforth referred to as the motion stimulus or the standard. The second object, ie the 'comparison', was defined by structure-from-retinal-disparity (Julesz 1971) and did not rotate. It was displayed until a response occurred for a maximum duration of 6 s. It is henceforth referred to as the stereo stimulus or the comparison. On match trials, the

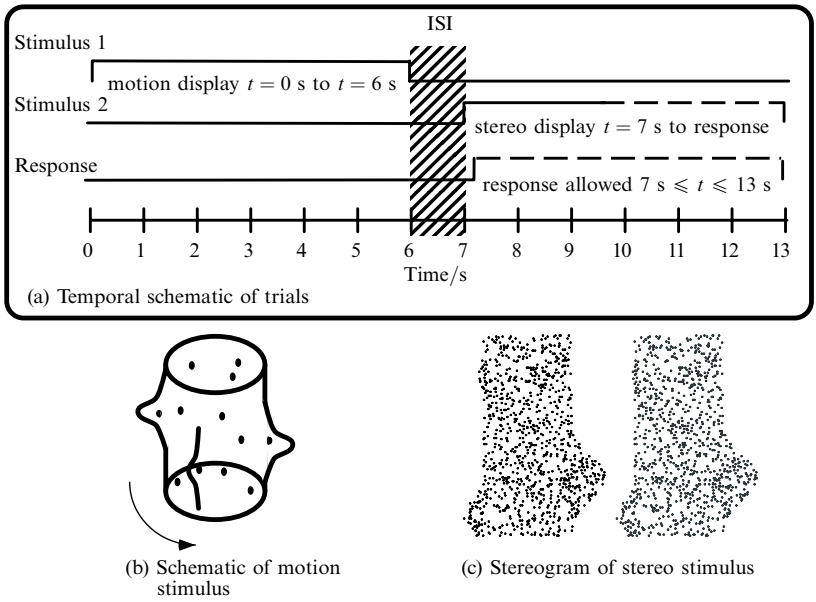

Figure 1. Trial timing and stimulus parameters. (a) Trials consisted of a 6 s motion stimulus, followed by a 1 s interstimulus interval (ISI) and up to 6 s of stereoscopic stimulus display. Responses terminated a trial and were allowed anytime after the second stimulus onset. If no response occurred during the second stimulus display, the trial was labeled 'incorrect' and tallied as either a miss (for a 'match' trial) or a false alarm (for a 'non-match' trial). After each trial, the performance was indicated by a tone, followed by a 1 s intertrial interval. (b) The motion-defined cylinder rotated 360° during its 6 s exposure duration. A parallel projection algorithm was used to compute dot positions. (c) The stereoscopic stimulus was stationary and oriented −60°, −30°, 0°, +30°, or +60° from the start/finish position of the motion stimulus.

stereo object appeared in depth offset by −60°, −30°, 0°, 30°, or 60° of rotation about the vertical axis relative to the final position of the motion object. Correct matches, false alarms, and reaction times were collected.

2.2 *Novel visual stimuli*

The form attributes of the objects were highly controlled by creating them from random dots of limited lifetime (533 ms or 32 display frames) with a constant point density (Morgan and Ward 1980; Siegel and Andersen 1988). The objects were selected from an infinite set of computer-generated transparent cylinders having three randomly placed Gaussian bumps on their surface (figures 1b and 1c). Bump placement could vary in height on the cylinder (h) and angular location on the cylinder surface (θ).

The bumpy stimuli were computed in cylindrical coordinates (R, θ, h) given in screen units (ie pixels), radians, and screen units, respectively. The radius at each point in the cylinder was given as $R(\theta, h) = C_r + \sum_{i=i}^{3} B_i(\theta, h)$. C_r was the radius of the cylinder (100 units). The Gaussian function $B_i(\theta, h)$ defining the ith bump was given as:

$$B_i(\theta, h) = A \exp\left[-\frac{(\theta - \theta_i)^2}{\lambda_\theta^2} - \frac{(h - h_i)^2}{\lambda_h^2}\right], \tag{1}$$

where $\lambda_\theta = \pi/9$ radians is the width of the Gaussian bump in θ, and $\lambda_h = 30$ is the amplitude in screen units. These two values were chosen to provide a roughly symmetrical bump when viewed end on. Three bumps were created by first setting the bump amplitude A to 100 units. The location of each bump was at a random location in cylindrical coordinates $\theta_i = [0, 2\pi]$ radians $h_i = [0, 200]$ and screen units. The amplitude of a bump

was thus equal to the radius of the cylinder (C_r). Bumps could overlap depending upon the values chosen for the three locations (θ_i, h_i), $i = 1, 2, 3$, resulting in larger bumps. All random number distributions were uniform. The displays were viewed at 57 cm so that 100 pixels was 2 deg.

To ensure subjects were not simply matching dot arrays, the two 'bumpy cylinders' were presented in different submodalities (either stereoscopic or structure-from-motion stimuli) with different arrays of random dots placed on their surfaces. Subjects reported a vivid impression of depth in both the motion and stereo stimuli.

Dot positions were computed by using a parallel lens-axis algorithm with perspective (Akka 1991). The viewpoint was placed at 57 cm with an interocular distance of 60 mm to compute the stereoscopic half-images. Stereoscopic viewing was accomplished with a Tektronix Stereoscopic Modulator (SGS610) running at 120 Hz (60 Hz to each eye) while observers wore left and right circularly polarized lenses on the left and right eyes, respectively. Each stimulus was composed of 1000 white dots of limited lifetime (533 ms) and 0.02 deg diameter on a black background. The dot patterns were thus dynamic over time and twinkled. The dots were randomly distributed on the surface of the display screen (not the surface of the object) so there were no density cues to shape. This was achieved through a search algorithm to first locate a point randomly placed on the monitor display and then back-project this point to one of many possible locations on the object. Stimuli were drawn from an infinite set of transparent cylinders 5.5 deg tall by 4.0 deg wide with three randomly placed Gaussian bumps on their surface.

2.3 *Experimental and control conditions*

Subjects performed the object recognition task in five different conditions. By manipulating the controlled visual display (eg using occluding contours), various hypotheses for the underlying mechanisms of visual recognition could be tested. In the normal view condition, all two-dimensional and three-dimensional shape cues were available for both the motion and the stereoscopic stimuli. This condition, termed Normal View, gave a baseline measure of subjects' ability to match the two stimuli. The two-dimensional cue to shape for the comparison stimulus was the outline or silhouette formed by the dots. The three-dimensional cue to shape was the retinal disparity gradient across the dots. These different shape cues were selectively deleted from particular conditions to allow for testing of subjects' abilities to use two-dimensional cues only, three-dimensional cues only, both two-dimensional and three-dimensional cues, or neither cue (Control condition). Since the Control condition (see below) required the use of occluders over the motion stimulus, all of the following conditions had occluders over the motion stimulus, to allow for better comparison across conditions.

A possible solution to the task in the Normal View condition is to match the stereo object's silhouette to a single 'temporal snapshot' of the motion object's silhouette by using two-dimensional cues. To prevent this from occurring, the contours formed at the edge of the dot displays were removed by placing two red opaque occluders over the edges of the displays in certain conditions. These occluders were 5.5 deg high $\times$ 2.0 deg wide (luminance 0.1 cd m^{-2}) and had an internal separation of 4 deg (the base diameter of the cylinders before bump placement). The occluders thus forced subjects to use only the motion flow or retinal disparity information derived from the 5.5 deg $\times$ 4 deg visible part of the display.

The condition in which the silhouette of the standard but not the comparison was occluded was termed 2-D + 3-D, to indicate that there were both two-dimensional cues and three-dimensional cues in the comparison stimulus. In this condition, subjects had to generate a representation of the first stimulus from motion cues only. However, both the silhouette and the disparity gradient in the stereoscopic stimulus could serve in representing its shape.

In order to remove the contribution of the two-dimensional contour shape information entirely, the outlines in the comparison display were occluded. Thus, only three-dimensional cues to shape, the retinal disparity gradient, were available in the comparison. Under this condition, termed 3-D-Only, the comparison display was a 5.5 deg $\times$ 4 deg rectangle of dots for which shape was defined only by disparity. In order to perform this task, subjects needed to extract a representation from the rectangle of motion flow in the occluded standard and then compare it to the retinal disparity in the occluded comparison stimulus.

In the 2-D-Only condition, both disparity and occluders were removed from the comparison stimulus. In this condition, the comparison stimulus contained only two-dimensional cues to object shape, ie the silhouette, since all disparities were set to zero. The silhouette was formed by the luminance boundary between the density of dots on the objects and the black background.

To demonstrate that uneven point density or other extraneous cues did not contaminate the stimuli, the Control condition was presented in which the second stimulus had no outlines and no disparity—presumably removing all cues for shape. This condition served as a control for both the 2-D-Only and 3-D-Only conditions. It is equivalent to the 2-D-Only condition with the outlines removed, or to the 3-D-Only condition with the disparity removed.

2.4 *Statistical analysis*

Hits, misses, false alarms, and correct rejections were recorded for each subject at each angle in each session. The hit rates and false alarms for each subject in each condition were submitted to a χ^2 ANOVA to determine whether subjects could reliably discriminate match from non-match trials (SAS procedure FREQ). Each subject in each condition failed to show an effect of angle. Once collapsed across angle, each subject's sensitivity (d') was computed for each condition.

3 Results

Prior studies have indicated that there is only a 2%–3% performance increase with the addition of three-dimensional information (Edelman and Bülthoff 1992; Blicher 1995; Bülthoff et al 1995). However, it was not considered whether three-dimensional cues alone might be just as useful for object recognition as two-dimensional cues alone. Thus, the working hypothesis was that if internal representations and the comparison process can be three-dimensional, then subjects should be able to perform the matching task even when they have access to only three-dimensional shape cues.

3.1 *Normal View condition*

The Normal View comparison utilized stimuli in which all two-dimensional and three-dimensional shape cues were present in both the standard and the comparison stimuli for each trial. All subjects were able to match the disparity objects to the motion objects regardless of relative orientation (figure 2). A χ^2 ANOVA showed no effect of angle upon the percentage of correct responses. Thus the data were collapsed across angles. Subjects S1 and S2 performed quite well using all the cues (S1, $d' = 3.5$; S2, $d' = 2.19$).

3.2 *2-D + 3-D condition*

In this condition the silhouette of the standard was occluded but not the comparison. Both two-dimensional cues and three-dimensional cues were left in the comparison stimulus. As in the Normal View condition, there was no significant dependence of the performance on angle for the 2-D + 3-D condition. Furthermore, the presence of the occluders over the standard did not alter either subject's ability to perform the matching task (S1, $d' = 3.93$; S2, $d' = 2.09$), suggesting that in the Normal View condition, subjects

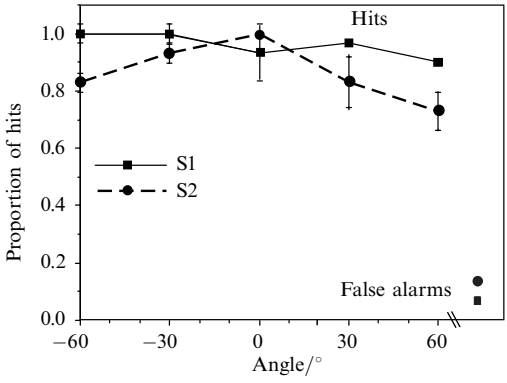

Figure 2. Hits and false alarms for the two subjects (S1 and S2) across all five angles. Each point represents the average taken from three sessions of 100 trials each (50 match, 50 non-match). A χ^2 test indicated that there was no effect of angle. Subjects could reliably discriminate 'match' from 'non-match' trials across all angles.

were indeed abstracting a surface representation based on motion flow, not on the stimulus silhouettes. This result suggests that subjects are able to obtain the three-dimensional shape of the first object using purely motion cues.

3.3 *3-D-Only condition*

Under this condition, subjects needed to extract a representation from the rectangle of motion flow in the occluded standard and then compare it to the retinal disparity in the occluded comparison stimulus. As in the other conditions, a χ^2 ANOVA indicated that there was no significant effect of the angle of the display upon performance. The data were combined across angles and it was found that the subjects performed well above chance in this condition (S1, $d' = 2.19$; S2, $d' = 1.14$), although not as well as in the 2-D + 3-D condition (figure 3). This experimental result is prima facie evidence that subjects can match objects in the absence of two-dimensional shape cues and is inconsistent with theories of object recognition that exclude three-dimensional stored representations.

3.4 *2-D-Only condition*

In this condition, the comparison stimulus was not masked and did not have any disparity cues. Thus it only contained two-dimensional cues to object shape (ie the silhouette). Subjects were able to match object shape in this condition. The χ^2 ANOVA showed no dependence of performance on angle. Performance in the 2-D-Only condition (S1, $d' = 2.00$; S2, $d' = 1.27$) was approximately the same as in the 3-D-Only condition but poorer than that in the 2-D + 3-D condition (figure 3). Thus, with the novel stimuli presented here, results were obtained which are consistent with previous claims that objects can be recognized on the basis of two-dimensional information alone (Poggio and Edelman 1990; Bülthoff and Edelman 1992; Edelman and Bülthoff 1992; Logothetis et al 1994; Blicher 1995; Bülthoff et al 1995; Logothetis and Pauls 1995; Vetter et al 1995).

3.5 *Control condition*

This condition served as a control for both the 2-D-Only and 3-D-Only conditions. The comparison stimulus did not have a silhouette owing to the occluders, and all disparities were zero. It was equivalent to the 2-D-Only condition with the outlines removed, or to the 3-D-Only condition with the disparity removed. Subjects performed at chance with this control (S1, $d' = -0.03$; S2, $d' = -0.05$), which suggests that there were no uncontrolled visual factors (eg dot density, or improper occluder placement) in either the 2-D-Only or 3-D-Only conditions.

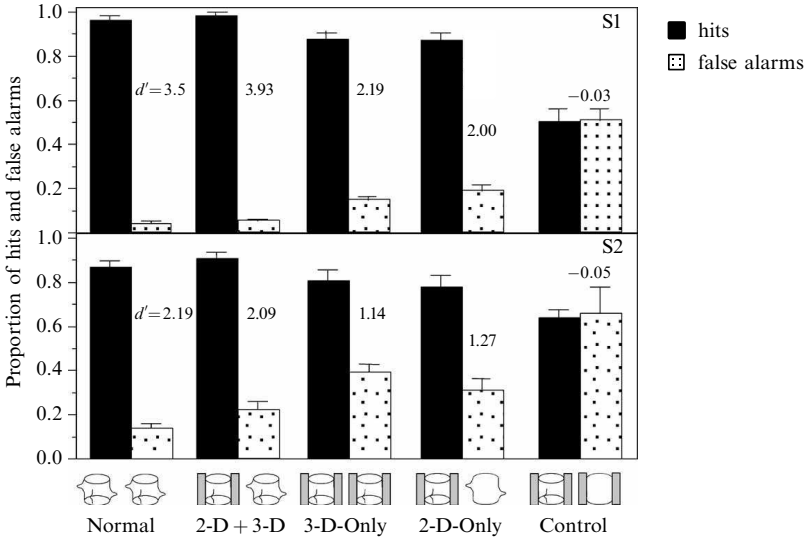

Figure 3. Hits, false alarms, and sensitivity (d') for S1 and S2 across the five viewing conditions. Subjects could reliably discriminate 'match' trials from 'non-match' trials in all but the Control condition. Notice that within subjects there is a very similar sensitivity in the 2-D-Only and 3-D-Only conditions, and slightly better sensitivity in the 2-D+3-D condition. This indicates that both two-dimensional and three-dimensional shape cues can be equivalently used to recognize objects. The similar results between the Normal View and 2-D+3-D conditions illustrate that the occluders on the motion stimulus had minimal, if any, effect on performance. The Control condition serves as both a two-dimensional cue control and a three-dimensional cue control. It is equivalent to removing disparity from the comparison in the 3-D-Only condition and to occluding the borders of the comparison in the 2-D-Only condition. Chance performance in this condition indicates that the appropriate cues were the only ones available in the 2-D-Only and 3-D-Only conditions.

4 Discussion

In the present study, two-dimensional and three-dimensional shape cues were independently assessed for their role in object recognition processes. This was possible through the use of displays that permitted independent control of each cue type. Subjects were able to abstract shape information from the motion displays and match it to that in the disparity displays even when only three-dimensional cues were available (3-D-Only condition). This result is inconsistent with theories that exclude storage of three-dimensional object representations. While the novel stimuli allowed an unprecedented control over shape cues, the specialized nature of the stimuli and viewing conditions limits the generality and scope of some of the conclusions. No statement as to viewpoint-invariant or viewpoint-dependent performance may be made because the stimuli were transparent and the standard was rotated through 360°, albeit quickly, providing a multiplicity of views.

4.1 *Choice of stimuli*

Stimulus choice provides strength to this study in that the 'bumpy cylinders' were visually similar, which is atypical in demonstrations of three-dimensional object recognition processes. All were cylinders of equal height and radius with three bumps on their surface that could vary in height (h), and angular location (θ). The objects appeared very similar to one another. Thus, a simple geon-structural description (Biederman 1987) would not suffice to produce the performance seen here, as subtle variations in bump height, angular placement, and overlap distinguish one cylinder from another. The conclusions are therefore not subject to the limitations of geon theory (see Tarr and Bülthoff 1995). Geon theory explains object recognition of different types of objects but

not recognition with a subtype of object. The demonstrated subordinate level recognition performance, based on small shape differences within a class of object (Rosch et al 1976), is inconsistent with claims that recognition at this level is mediated only by two-dimensional, view-based information (Logothetis et al 1994).

4.2 *The necessity of three-dimensional representations*

The information in the motion displays was derived from different speeds of moving dots. These dots moved to define a three-dimensional object and were based upon a rich history of psychophysical studies starting with the kinetic depth effect (Wallach and O'Connell 1953). The occluded display clearly looked as if one were viewing a rotating three-dimensional object through a window; however, it is an assertion that the subjects necessarily represented the objects as three-dimensional. Presumably the motion gradients define the three-dimensional surfaces.

Surprisingly, a similar statement can be made for the stereoscopic comparison display. It clearly looks three-dimensional, and indeed the disparities define three-dimensional surfaces much as the motion gradient does in the standard motion display. However, there is a difficult problem in that neither the occluded motion display nor the occluded stereoscopic display absolutely defines a three-dimensional shape. There is the possibility, however unlikely, that the subjects are matching regions of high dot speed to regions of high dot disparity. This could then lead to the suggestion that the task was done through matching of two-dimensional cues.

One result strongly argues against this interpretation of the matching paradigm. One would expect that, if the subjects were to match a particular region of high speed to high disparity, then spatially separating the two stimuli by changing the angle of rotation should make the two difficult to match for some angle. Performance does not appear to depend on the angle of presentation of the second display relative to the first display. To press this two-dimensional disparity/motion matching explanation even further, one could argue that the first display has presented all the angles; thus all the subject needs to do is choose the moment of the standard motion display that matches the proper region of high disparity in the stereo display. However, this should be quite difficult considering that the motion standard precedes the disparity display. The subject would need to have a complete memory of the motion display.

Another possible strategy would be simply to match the locations of the bumps (h_i) along the vertical dimension. During the rotating motion display, points are moving to the left and right at speeds determined by their three-dimensional coordinates and the rate of rotation. Bumps by definition will result in larger radii and thus higher speeds. This possibility, to choose a solution based simply on the height of the bumps, could work if the precise value for each of the bumps can be computed from the velocity distribution.

This possibility was tested explicitly by generating a set of masked displays in which the distribution of speeds was maintained as a function of height, but the horizontal distribution of speeds was disrupted. Thus if subjects were solely using the vertical distribution of speeds to determine where bumps were located, then they would be able to match the high-speed regions with the presence of bumps in the stereo display. The disruption of the vertical distribution of speeds was performed by the 'unstructured motion' technique introduced by Siegel and Andersen (1988). The position of each motion trajectory was randomly displaced horizontally within a window of 100 screen units. The vertical position of each trajectory was unchanged. Note this unstructuring also degrades the smoothness of the speed profile, which may result in some disturbance to the two-dimensional percept.

Three subjects were given the task of matching the unstructured to the structured displays. Contours were masked in both conditions. As in Siegel and Andersen (1988), three subjects were able to differentiate the unstructured motion displays from the structured display. However, the three subjects were unable to match the objects defined by the unstructured displays to those of the structured displays by visual inspection.

The reason for the inability to extract the height of the bumps from the speed distribution is that, while the unstructured displays contain the exact same vertical gradients of motion, the horizontal gradients, which are necessary for the extraction of three-dimensional shape, are incorrect. It is only possible to extract the bump location given the complete three-dimensional information, which can be derived from the vertical and horizontal speed gradients.

These gradients were specifically examined. A typical object of 270 frames with 1000 points per frame was computed. The horizontal speed was computed as the difference in position between a point's location in two subsequent frames (Siegel and Read 1997). The number of points with a particular combination of horizontal speeds and vertical position was computed for all 270 frames (figure 4b). It can be seen that across all vertical positions there were two invariant peaks in the distribution corresponding roughly to ± 5 deg s^{-1}. The dependence of the height of the peaks at approximately -5 deg s^{-1} is illustrated in figure 4a (yellow line). These peaks arise from the rotation of the cylinder with radius C_r. The distribution of higher and lower speeds (ca ± 8 deg s^{-1}) varied with height (figure 4a, red line). This variation with height of a small number of points was a result of the bumps with their larger radii. Exactly the same distribution was seen with the unstructured and structured motion.

Thus, although there is a similar gradient with vertical position for the structured and unstructured motion, subjects are only able to extract the bumps in the display when the complete horizontal and vertical motion gradients are available. As these two gradients define the three-dimensional location of the bumps (Longuet-Higgins and Prazdny 1980), it seems most likely that the subjects are locating the bumps using motion cues that define a three-dimensional location, and then use these locations to make the match to the disparity-defined object. Ultimately, however, one must acknowledge that there is an intrinsic and profound difficulty in separating out the three-dimensional percept from the two-dimensional spatial speed gradient or from the two-dimensional distribution of disparity. When the visual system is confronted with this same problem, it may find that it is best to use both the three-dimensional and two-dimensional representations of the external world.

Thus, it is concluded that to perform in the condition for which three-dimensional shape cues could only be obtained from motion or depth, subjects were extracting three-dimensional information. The three-dimensional information was extracted from the motion display and matched to that of the three-dimensional surfaces derived from disparity. Control experiments show that the point density in the displays could not account for the performance. Further, subjects could not do this task by simply matching the velocity gradient in the vertical dimension with the location of a bump defined from motion or disparity.

4.3 *The necessity of two-dimensional representations*

The necessity for a three-dimensional representation does not obviate the need for a two-dimensional representation. Indeed, in our paradigm, subjects were able to perform equally well using two-dimensional cues only. This would fit well with the published literature (Edelman and Bülthoff 1992; Blicher 1995; Bülthoff et al 1995). As in these other papers, a slight increase in performance was seen when both two-dimensional and three-dimensional information was present.

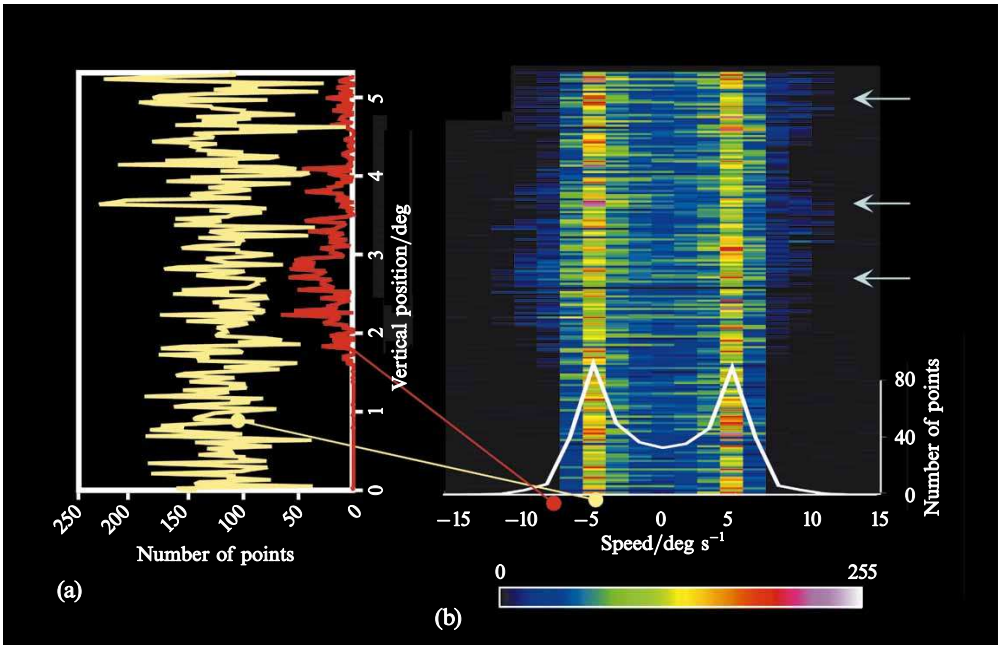

Figure 4. Distribution of horizontal speeds in the bumpy-object motion displays. In order to compute the speed for each point (j) in every frame (i) of the display the equation $V_{jx}^{i} = (P_{jx}^{i+1} - P_{jx}^{j})$ was used with appropriate scaling for time. A bumpy object with bumps at locations $(\theta_i, h_i) = (-0.05$ rad, 5 deg), (1.55 rad, 2.8 deg), (0.47 rad, 3.7 deg) was generated. Displays of 270 frames with 1000 points per frame were generated. The display rotated at 60 deg s^{-1} about the vertical axis. The heights of the three bumps are indicated by the blue arrows at the right of panel (b). (a) The number of points with a speed of -4.9 deg s^{-1} (yellow line) as a function of height is shown. There is substantial variation about the mean of 118 points with no particular dependence on the height of the bump. The number of points with a speed of -8.2 deg s^{-1} as a function of height (red line) is also shown; at this more negative speed the dependence on the height is more apparent in the graph. However, the number of points at this speed is about 10% that at the speed of -4.9 deg s^{-1}. The cumulative sums over 270 frames are shown. (b) The complete distribution of the number of points as a function of height and horizontal speed is presented. The color bar indicates the scaling of the number of points accumulated over 270 frames. The relative invariance of the horizontal speeds as a function of height may be seen for the slower speeds. At the higher absolute values of speed, there is a systematic variation of the speeds with height roughly corresponding to the location of the bumps. However, the numbers of points exhibiting this variation are fewer. The white line that is overlaid on panel (b) illustrates the distribution of speeds taken across all heights.

The present results indicate that human subjects are not limited to using only two-dimensional representations (Bülthoff et al 1995), but rather may use flexible representations which can include detailed depth information. Viewpoint-dependent performance in human observers does not require two-dimensional view-based storage in memory (Liu 1996). The present study further questions the sufficiency of two-dimensional view-based information in the object recognition process. Indeed, our result that subjects do indeed use three-dimensional cues agrees with those of Liu et al (1995) which demonstrate that human observers were more accurate than an ideal observer model which utilized only two-dimensional view-based information for object recognition.

These findings indicate that internal representations of objects can be three-dimensional in nature. This implies that the comparator itself is likely a three-dimensional process; not in the sense of requiring depth information but in the sense of being able to process it (when present) and use either two-dimensional or three-dimensional

information alone if warranted. Most current view-based models of object recognition must be generalized in order to incorporate three-dimensional object representation.

4.4 *Implications for neurophysiological studies*

These issues impact physiological studies that explicitly examine internal representations by neurons. The comparison of the disparity-defined and the motion-defined objects should occur in cortical regions where the two submodalities converge. The earlier regions of convergence (MT/V5, MST—see Maunsell and Van Essen 1983; Roy et al 1992) have smaller receptive fields and thus would be less likely to subserve the three-dimensional matching process. They could serve to indicate when regions of high two-dimensional motion flow matched regions of high disparity. However, it is not clear how these cortical areas could make the transformations needed to compare the motion and disparity signals when they are not spatially (ie retinotopically) superimposable. Other candidates would be 7a, STPa, and IT. Area 7a combines complex optic flow, extraretinal information, and disparity (Sakata et al 1980; Phinney and Siegel 1997, 1998; Read and Siegel 1997). Both 7a and MST project into STPa, which has neurons that exhibit structure-from-motion selectivity (Bruce et al 1981; Anderson and Siegel 1997).

In the ventral stream, recordings in IT have indicated that there are neurons that represent objects in three-dimensional, object-based coordinates on the basis of simple stimuli (eg Schwartz et al 1983; Perrett et al 1991). More careful studies that specifically tested for object-centered versus view-based effects suggested that the majority of neurons were in two-dimensional, view-based coordinates although a small percentage were found in temporal lobe, which appeared to clearly have a three-dimensional representation (Logothetis and Pauls 1995). While the majority of IT neurons appear to encode viewpoint-dependent information, some cells do exhibit viewpoint-invariant responses to objects. The predominantly two-dimensional viewpoint-dependent performance of these IT cells has been used to argue in favor of two-dimensional view-based object recognition (Logothetis et al 1994). However, an alternative explanation is that the small minority of cells which have been classified as view-invariant may be the cells which subserve object recognition decisions, whereas the viewpoint-dependent cells might be considered the primitive neural representation from which the view-invariant cells draw their information. Additional studies in the temporal lobe indicate the presence of neurons selective to the three-dimensional orientation of objects defined by disparity (Janssen et al 1997), while studies in the parietal lobe show neurons clearly working in an object-centered coordinate system during grasping paradigms (Sakata et al 1995). In these physiology studies, as in the psychophysical ones, it is difficult to resolve unambiguously the contribution of two-dimensional and three-dimensional cues.

5 Conclusion

Human subjects reliably matched object shapes when the objects contained only three-dimensional visual form cues. Similar performance was found when objects contained only two-dimensional cues. The parity in performance for these two cues indicates that both two-dimensional and three-dimensional information can effectively support object recognition. There was only a small increase in performance when both two-dimensional and three-dimensional shape cues were present, as might be expected for highly redundant information. These findings indicate that three-dimensional information can be used to recognize objects. Further implied is that object recognition processes utilize three-dimensional stored representation in the absence of, and in addition to, two-dimensional representations. This requires that any theory of human object recognition, object-centered or view-based, allows for the use of three-dimensional stored representation.

Acknowledgements. We acknowledge Carlos A M Nogueira for programming the display software and for critical help in establishing parameters in the pilot work. This work was supported by the Office of Naval Research Grant Number N00014-93-1-0334 and NIH EY-0992. RP was supported by NIH EY06738-01.

References

Akka R, 1991 "Creating stereoscopic software", in *The Crystal Eyes Handbook* Ed. S Boris (San Rafael, CA: Stereographic Corporation) pp 31 – 41

Anderson K C, Siegel R M, 1997 "Distribution of optic flow selectivities in the anterior superior temporal polysensory area (STPa) in the behaving macaque" *Society for Neuroscience Abstracts* **22** 460

Biederman I, 1987 "Recognition-by-components: a theory of human image understanding" *Psychological Review* **94** 115 – 147

Blicher A P, 1995 "A shape representation for computer vision based on differential topology" *Biosystems* **34** 197 – 224

Bruce C, Desimone R, Gross C G, 1981 "Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque" *Journal of Neurophysiology* **46** 369 – 384

Bülthoff H H, Edelman S, 1992 "Psychophysical support for a two-dimensional view interpolation theory of object recognition" *Proceedings of the National Academy of Sciences of the USA* **89** 60 – 64

Bülthoff H H, Edelman S Y, Tarr M J, 1995 "How are three-dimensional objects represented in the brain?" *Cerebral Cortex* **5** 247 – 260

Edelman S, Bülthoff H H, 1992 "Orientation dependence in the recognition of familiar and novel views of three-dimensional objects" *Vision Research* **32** 2385 – 2400

Janssen P, Vogels R, Orban G A, 1997 "Responses of monkey inferior temporal neurons to disparity gradients" *Society for Neuroscience Abstracts* **23** 2065

Julesz B, 1971 *Foundations of Cyclopean Perception* (Chicago, IL: University of Chicago Press)

Liu Z, 1996 "Viewpoint dependency in object classification and recognition" *Spatial Vision* **9** 491 – 521

Liu Z, Knill D C, Kersten D, 1995 "Object recognition for human and ideal observers" *Vision Research* **35** 549 – 568

Logothetis N K, Pauls J, 1995 "Psychophysical and physiological evidence for viewer-centered object representations in the primate" *Cerebral Cortex* **3** 270 – 288

Logothetis N K, Pauls J, Bülthoff H H, Poggio T, 1994 "View-dependent object recognition by monkeys" *Current Biology* **4** 401 – 414

Longuet-Higgins H C, Prazdny K, 1980 "The interpretation of a moving retinal image" *Proceedings of the Royal Society of London, Series B: Biological Sciences* **208** 385 – 397

Marr D, Nishihara H K, 1982 "Representation and recognition of the spatial organization of three-dimensional shapes" *Proceedings of the Royal Society of London, Series B: Biological Sciences* **2** 269 – 294

Maunsell J H, Van Essen D C, 1983 'The connections of the middle temporal visual area (MT) and their relationship to cortical hierarchy in the macaque monkey" *Journal of Neuroscience* **3** 2563 – 2586

Morgan M J, Ward R, 1980 "Conditions for motion flow in dynamic visual noise" *Vision Research* **20** 431 – 435

Perrett D, Oram M W, Harries M, Bevan R, Hietanen J K, Benson P J, Thomas S, 1991 "Viewer-centered and object-centered coding of heads in macaque temporal cortex" *Experimental Brain Research* **86** 159 – 173

Phinney R E, Siegel R M, 1997 "Cells in macaque area 7a are sensitive to retinal disparities embedded in optic flow stimuli" *Society for Neuroscience Abstracts* **23** 1546

Phinney R E, Siegel R M, 1998 "Disparity selectivity and optic flow in monkey area 7a" *Society for Neuroscience Abstracts* **24** 1141

Poggio T, Edelman S, 1990 "A network that learns to recognize three-dimensional objects" *Nature (London)* **343** 263 – 266

Read H L, Siegel R M, 1997 "Modulation of responses to optic flow in area 7a by retinotopic and oculomotor cues in monkey" *Cerebral Cortex* **7** 647 – 661

Rosch E, Mervis C B, Gray W D, Johnson D M, Boyes-Braem P, 1976 "Basic objects in natural categories" *Cognitive Psychology* **8** 382 – 439

Roy J P, Komatsu H, Wurtz R H, 1992 "Disparity sensitivity of neurons in monkey extrastriate area MST" *Journal of Neuroscience* **12** 2478 – 2492

Sakata H, Shibutani H, Kawano K, 1980 "Spatial properties of visual fixation neurons in posterior parietal association cortex of the monkey" *Journal of Neurophysiology* **43** 1654 – 1672

Sakata H, Taira M, Murata A, Mine S, 1995 "Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey" *Cerebral Cortex* **5** 429 – 438

Schwartz E L, Desimone R, Albright T D, Gross C G, 1983 "Shape recognition and inferior temporal neurons" *Proceedings of the National Academy of Sciences of the USA* **80** 5776 – 5778

Shepard R N, Metzler J, 1971 "Mental rotation of three-dimensional objects" *Science* **171** 701 – 703

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

Siegel R M, Read H L, 1997 "Analysis of optic flow in the monkey parietal area 7a" *Cerebral Cortex* **7** 327 – 346

Tarr M J, Bülthoff H H, 1995 "Is human object recognition performance better described by geon-structural-descriptions or multiple views?" *Journal of Experimental Psychology: Human Perception and Performance* **21** 1494 – 1505

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

Ullman S, 1989 "Aligning pictorial descriptions: an approach to object recognition" *Cognition* **32** 193 – 254

Vetter T, Hurlbert A, Poggio T, 1995 "View-based models of 3D object recognition: Invariance to imaging transformations" *Cerebral Cortex* **5** 261 – 269

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