Recognizing and segmenting objects in clutter

MARY J. BRAVO, [1] HANY FARID [2]

When viewing a cluttered scene, observers may not be able to segment whole objects prior to recognition. Instead, they may segment and recognize these objects in a piecemeal way. Here we test whether observers can use the appearance of one object part to predict the location and appearance of other object parts. During several training sessions, observers studied an object against a blank background. They then viewed this object against a background of clutter that camouflaged some parts of the object while leaving other parts salient. The observer's task was to find the camouflaged part. We varied the symmetry of the salient part with the expectation that as this symmetry decreased, the information about the camouflaged part's location and appearance would increase and this would facilitate search. Our results suggest that observers can use the salient part to predict the location, but not the appearance, of the camouflaged part.

Object Segmentation Object Recognition

[1]Department of Psychology, Rutgers University, Camden NJ 08102. Email: mbravo@camden.rutgers.edu; Tel: 856.225.6431; Fax: 856.225.6602

[2]Department of Computer Science and Center for Cognitive Neuroscience, Dartmouth College, Hanover NH 03755.

1 Introduction

In a cluttered scene it may not be possible to segment objects without assistance from top-down processes (Barrow & Tenenbaum, 1981; Marr, 1982; Spelke, 1990; Ullman, 1997; Borenstein & Ullman, 2002; Bravo & Farid, 2003). Consider Figure 1 which shows a scene composed of several familiar objects with similar colors and textures. Note that because the objects occlude one another, parts from different objects are intermingled. If bottom-up grouping processes alone were used to organize this scene, it seems unlikely that the result would correspond to the familiar objects that we perceive.

Although researchers in human vision have largely ignored the problems of segmenting and recognizing objects in clutter (c.f., (Brady & Kersten, 2003)), a number of researchers in computer vision have proposed approaches to this problem (Roberts, 1966; Lowe, 1985; Selinger & Nelson, 1999; Borenstein & Ullman, 2002). One such approach, sometimes termed "hypothesize-verify", involves the following steps. Simple grouping processes organize the image into parts (whether these parts correspond to contours or image patches or volumes depends on the model). Some parts are so distinctive that they may be recognized as likely belonging to a particular object or class of objects. Recognition of such a part allows the observer to form a hypothesis about an object in the scene, and this hypothesis allows the observer to predict other parts of the object that should be present in the image. If these parts are found in the image, then the hypothesis is verified, and the object is recognized. Applying this scenario to Figure 1, the observer might first tentatively recognize the roll of tape (A). Then, suspecting that a tape dispenser appears in the scene, the observer may look for other parts of this object, such as the end with the serrated blade (B).

For a part to be useful in generating a hypothesis, it should be readily extracted from most images that contain the object. In other words, the part should generally be present in the image when the object is present in the scene. Further, the part should be diagnostic of the object. That is, the part should be absent from the image when the object is absent from the scene. Because the first requirement favors simple parts while the second requirement favors complex parts, the most useful parts are likely to be those of intermediate complexity (Ullman, Sali, & Vidal-Naquet, 2001).

After a distinctive part has been recognized and the observer has formed a hypothesis about the object in the scene, the observer must then verify this hypothesis. As noted above, this verification involves predicting other object parts that should exist in the scene and then either confirming that these parts do indeed exist or determining that their absence can be explained by an occlusion. In this verification process, it is possible that the observer could use the *appearance* of the recognized part to predict both the *location* and the *appearance* of the other object parts. That is, through repeated exposure to different views of the object, the observer may learn to associate, for each view, the spatial relationships and appearances of various object parts. Then, given the appearance of one part, the observer may recall the location and appearance of other object parts from the same view. Alternatively, the observers may form what might be called a "cubist" representation of the object. [3] That is, they may associate the various appearances of the object parts in a way that is largely independent of viewpoint. With this kind of representation, a distinctive part would cue all of the part appearances that are associated with the object, not just those from the same viewpoint.

[3]The idea that our internal representation of objects might have certain resemblances to the cubist paintings of Picasso and Braque was proposed by Nelson and Selinger (Nelson & Selinger, 1998).

Figure 1: This scene contains a partially occluded tape dispenser (A and B). It seems unlikely that bottom-up grouping processes would associate the image fragments that correspond to A and B.

In the experiments reported here, we examined the verification stage of the hypothesize-verify approach to object recognition. We assumed that the appearances of object parts are stored in a viewpoint specific way. Thus, when observers verify that an object is in the scene, they can use appearance of a distinctive part to predict the location and the appearance of other object parts. From this we reasoned that distinctive parts with zero rotational symmetry should be particularly useful for making such predictions. This is because each appearance of the part is associated with a particular object pose. Thus these parts make a single, reliable prediction. Distinctive parts that have non-zero rotational symmetry should be less useful for finding the rest of the object. For these parts, one appearance is associated with multiple object poses. Thus these parts will make multiple predictions, all but one of which will be wrong. In general, we expected that performance should be inversely proportional to the rotational symmetry of the distinctive part.

We tested this idea in three experiments. In these experiments, we presented an object in a random pose against a background of clutter. Because this background clutter closely resembled some parts of the object, it camouflaged these parts, while leaving other parts quite salient. We examined whether the symmetry of the salient part affected the observer's ability to find the camouflaged part. In all experiments, the pose of the object was varied across trials: in the first experiment, the objects were rotated in the image plane, while in the second and third experiments, the objects were rotated in depth.

2 Experiment 1: Rotations in the Image Plane

In this first experiment, in which we confined object rotations to the image plane, we used the four objects shown in Figure 2. Each object consisted of a geometric shape (triangle, rectangle, square, or circle) and two cylinders, which we will refer to as a handle. During two training sessions, observers learned these objects by viewing them against a blank background. Then during the testing session, observers viewed the objects against a background of cylinders that were identical

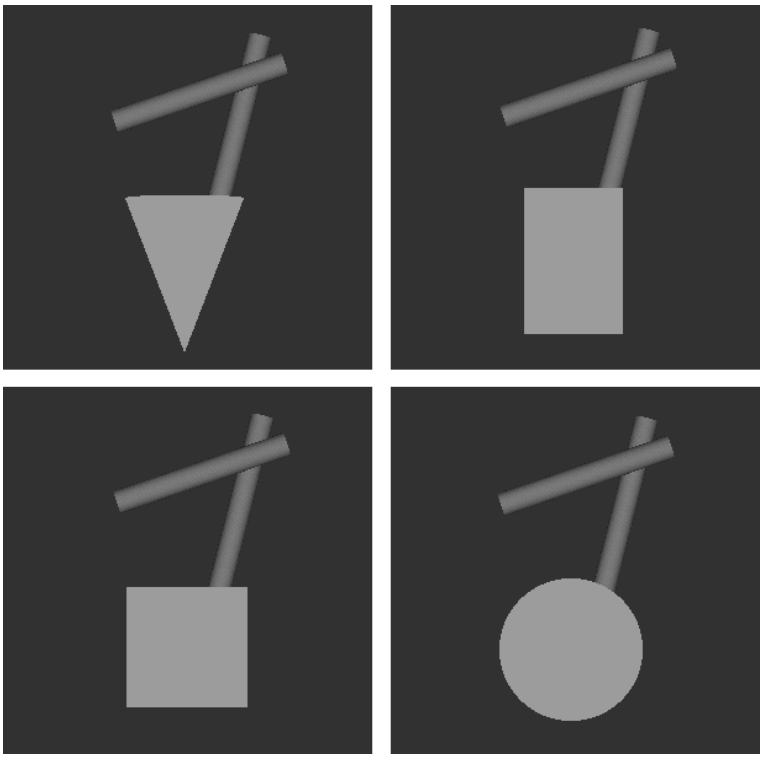

Figure 2: Objects used in Experiment 1. Each object consisted of two cylinders (the handle) and a geometric shape. The geometric shapes differed in their degree of rotational symmetry.

to those composing the handle. Against this background, the handle was well-camouflaged, but the geometric part was salient, Figure 3.

We placed a black or white ring in a random location on either cylinder of the handle. The observer's task was to report the color of this ring. Because rings were also placed on the clutter, accurate performance required that observers find the handle. We were interested in how the symmetry of the geometric shape would affect search for the handle. We expected that the shape with the lowest rotational symmetry, the isosceles triangle, would provide an unambiguous prediction of the handle's angular location and so would produce the fastest responses. In contrast, the shape with perfect rotational symmetry, the circle, would provide no information about the handle's angular location, and so would produce the slowest responses.

2.1 Stimuli

The stimuli were rendered as three dimensional objects using OpenGL on a Silicon Graphics O2. For accurate response timing, stimuli were displayed on an Apple PowerBook using MatLab and PsychToolbox routines (Brainard, 1997; Pelli, 1997).

Each object was composed of two parts: a geometric shape and two cylinders, Figure 2. The radius of the circle was 1.75 degrees of visual angle from the approximate viewing distance of 60 cm. The other shapes were designed to have the same perimeter as the circle. We matched the perimeters so that when the shapes were presented against the cluttered background (see below) each shape's boundary would contact a similar amount of clutter.

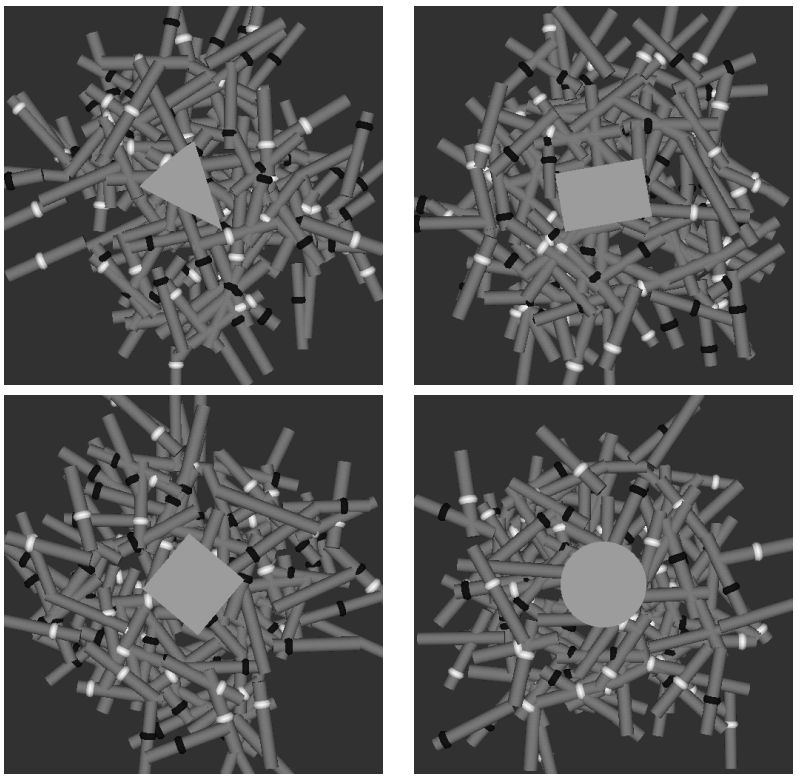

Figure 3: The objects from Experiment 1 presented against a background that camouflaged the handle. The observers' task was to find the handle (see Figure 2) and report whether it had a black or white ring.

During training, the objects were presented against a dark gray background. During testing, these objects were presented against a background of 100 cylinders, each with the same length and width as the cylinders forming the handle, Figure 3. These background cylinders were presented at random locations and orientations within a circular region with a radius of 12 degrees of visual angle. The background cylinders were placed in a slightly lower depth plane so that they would not occlude or penetrate the handle. The background cylinder locations and orientations were re-randomized on each trial. Also during testing, a black or white ring was placed in a random location on one of the cylinders composing the handle. Half of the background cylinders were also given a randomly positioned black or white ring. The observer's task was to find the handle and report the color of its ring.

2.2 Procedure

2.2.1 Training

Each observer participated in two training sessions and one testing session. These sessions occured on three separate days during a one week interval. During the training sessions, observers viewed a sequence of images of each object. The object was shown at 18 orientations, presented in

random order. These orientations sampled the full range of image rotations in intervals of 20 degrees. A sequence consisted of 36 images presented for five seconds each. During the two training sessions, observers watched six sequences for each of the four objects, resulting in 18 minutes of exposure for each object. In addition to passively viewing these images, observers also practiced making the appropriate response to the ring colors.

2.2.2 Testing

After the two training sessions, observers returned for a testing session in which they saw the objects against a background of clutter that camouflaged the handle. The observer's task was to report whether the ring on the handle was black or white. During this testing session, each of the four objects was presented in four separate blocks of trials. The order of these blocks was balanced across observers.

A block of trials began when the observer initiated the first stimulus presentation. The stimuli were presented until the observer responded, and if the response was incorrect, auditory feedback was given. Between stimuli, there was a 250 msec blank interval. Each block consisted of 30 trials, but the first two trials were discarded.

2.3 Participants

All of our participants were undergraduate students at Rutgers-Camden. They participated for class credit or for pay. None of the observers participated in more than one experiment, and none were informed of the specific predictions of these experiments.

2.4 Results and Discussion

The response times for the four observers are shown in Figure 4(a). The horizontal axis indicates the order of symmetry for the various geometric shapes. Recall that the rotational symmetry of the shape corresponds to the number of predictions that it makes about the handle's location and appearance. The vertical axis shows the time required for observers to correctly identify the color of the ring on the object's handle. An ANOVA indicated that there was a significant effect of the rotational symmetry on response times, $F(3,12) = 7.591$, $p = 0.004$. As expected, the triangle produced the fastest responses. This shape has a rotational symmetry of order one and so makes a single prediction about the handle. The rectangle, which has rotational symmetry of order 2, produced slower responses. The square, which has rotational symmetry of order 4, produced still slower response. Interestingly, the circle, which has infinite rotational symmetry and so made no predictions about the handle, produced reaction times that were slightly faster than those for the square (paired t-test: $t = 6.5109$, $p = 0.007$). [4]

The average response accuracy of the four observers is shown in Figure 4(b). The horizontal axis is the same as in the response time graph. These accuracy data indicate that the response time effects were not due to a speed-accuracy trade-off: accuracy levels were similar across object types.

[4]When we ran this experiment with mixed blocks, we also found that performance declined as the cue's order of symmetry increased up to four. For one observer this decline was reflected in an increased reaction time similar to the pattern seen in Figure 4. For two observers, however, the decline was reflected in an increased error rate. Thus, the blocked design of this experiment may be important for obtaining a consistent effect on response time.

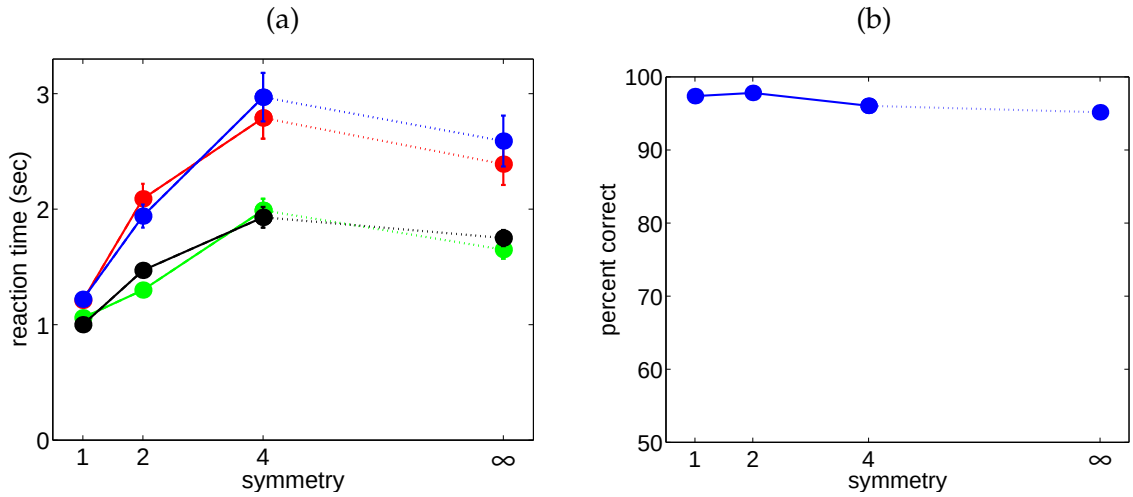

Figure 4: Results from Experiment 1. (a) Reaction times to find the camouflaged object part plotted against the rotational symmetry of the salient object part. (b) Accuracy averaged across the four observers.

We had expected that response times would increase as the rotational symmetry of the salient part increased. And while this was true when the number of predictions is relatively small, it was not true when the number of predictions exceeds four. For all observers, response times for the circle were faster than those for the square, even though the circle provided no angular information about the handle's angular location and appearance. Because the circle provided no such information, observers presumably used a global search strategy to find the handle of this object. That is, instead of directing their attention to specific locations, they may have simply searched for some distinctive feature of the handle. After running the experiment, several observers reported that one such distinctive feature was the orientation of the most central cylinder of the handle. This cylinder was oriented roughly along a line radiating from the center of the stimulus. Apparently, searching globally for a radially oriented cylinder was more efficient than searching selectively in four locations.

To return to the question motivating this experiment, we did find evidence that the appearance of the salient, geometric shape was used to make a prediction about the camouflaged handle. It is not clear from this experiment, however, what sort of prediction observers were making. Observers may have predicted the handle's appearance as well as its location. For example, when the triangle pointed up, observers may have searched below the triangle for two cylinders forming an "L" shape, and when the triangle pointed down, observers may have looked above the triangle for two cylinders forming an upside down "L". Alternatively, the observers may have only predicted the location of the handle. That is, they may have simply looked for any cylinder that was adjacent to a particular spot on the base of the triangle. To try to dissociate this kind of spatial cuing from appearance cuing, we repeated this experiment with objects that rotated in depth. We designed these experiments so that the cue would predict the appearance but not the location of the target.

3 Experiment 2: Rotations in Depth

This second experiment resembled the previous experiment in that observers were presented with an object against a background that effectively camouflaged the handle of the object while leaving other object parts salient. Once again, we were interested in how the symmetry of the salient part would affect the observer's ability to find the camouflaged handle. In contrast to the previous experiment, the objects in this experiment rotated in depth.

Two versions of this experiment were conducted. In one, observers were trained and tested on only three views of the object. These three views spanned a 90 degree rotation and corresponded to three very distinct appearances. In the other version of the experiment, different observers were trained and tested on 31 views that spanned the same 90 degree range. Thus, in this second version of the experiment, observers were exposed to a continuum of object appearances.

3.1 Stimuli

Three views of the cone object are shown in Figure 5. This object consisted of a long vertical pole with a cone at the top and four connected cylinders near the bottom. [5] To be consistent with the previous experiment, we will refer to these four cylinders as the handle. Note that as the object rotated, the location of the handle remained fairly constant even though its appearance changed markedly. Because the cone had an axis of revolution that was orthogonal to the axis of object rotation, its appearance changed as well. We also generated a control object, which had a large cylinder in place of the cone, but was otherwise the same, Figure 6(b). Unlike the cone, the cylinder had an axis of revolution that coincided with the axis of rotation. As a result, the cylinder's appearance never changed.

Although the handle and the cone (or cylinder) rotated rigidly, we allowed the handle to slide up and down the vertical pole as in Figure 6(a). We added this random translation because in pilot studies we were unable to get an effect of appearance cuing without some uncertainty about the location of the handle. We also randomly shifted the horizontal location of the whole object. These horizontal shifts were intended to discourage observers from using absolute spatial cues to locate the handle.

During the training phase, the objects were presented against a medium gray background. During the testing phase, the objects were presented against a background of clutter that filled the 24° by 24° display, Figure 7. This background clutter consisted of about 600 cylinders with lengths and orientations that matched those of the cylinders composing the handle. In the test stimuli, a black or white ring was positioned in a random location on one of the four cylinders of the handle. One quarter of the background cylinders were also given a black or white ring. In this experiment, the object and background clutter were rendered separately, and then the object was superimposed on the background. To eliminate image-processing artifacts, the two images were combined using anti-aliasing techniques.

[5] The range of orientations was limited to those in which the handle remained behind the vertical pole. If the handle had occluded the pole, this would have provided a salient cue to its location.

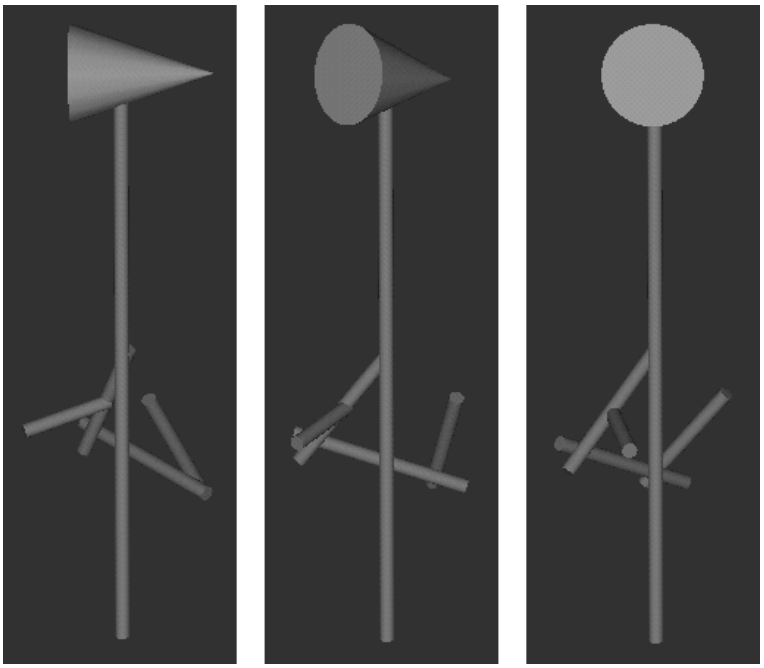

Figure 5: The cone object used in Experiment 2. The object consisted of four cylinders (the handle) connected by a long pole to a cone. As this object rotated about the vertical axis, the appearance of the handle changed.

3.2 Procedure

3.2.1 Training

The training procedure was very similar to that used in the previous experiment. Observers viewed sequences of images showing different views of the objects. For the discrete (three-view) version of the experiment, observers had two training sessions and saw the cone and cylinder objects each for a total of 18 minutes. For the continuous (31-view) version of the experiment, we assumed that observers would need more training to learn the additional views. Observers had four training sessions and saw the cone and cylinder objects each for a total of 36 minutes.

3.2.2 Testing

During the testing phase we presented the objects against a background of clutter that camouflaged the handle but left the geometric shape salient, Figure 7. As in the previous experiment, the observer's task was to report the color of the ring on the handle. In contrast to the previous experiment, however, the observers knew the handle would appear in the bottom part of the display. Because we were concerned that observers might attend only to this region, we briefly occluded the bottom part of the display with a gray rectangle, leaving the geometric shape visible. After 500 msec, this rectangular occluder was removed, and the whole stimulus was visible. Response times were measured from the offset of the occluder. The response times for the first two trials of each block were discarded.

During the testing session, the cone and cylinder objects were presented in separate blocks of trials. Five blocks, each consisting of 30 experimental trials, were run for each object. The order of

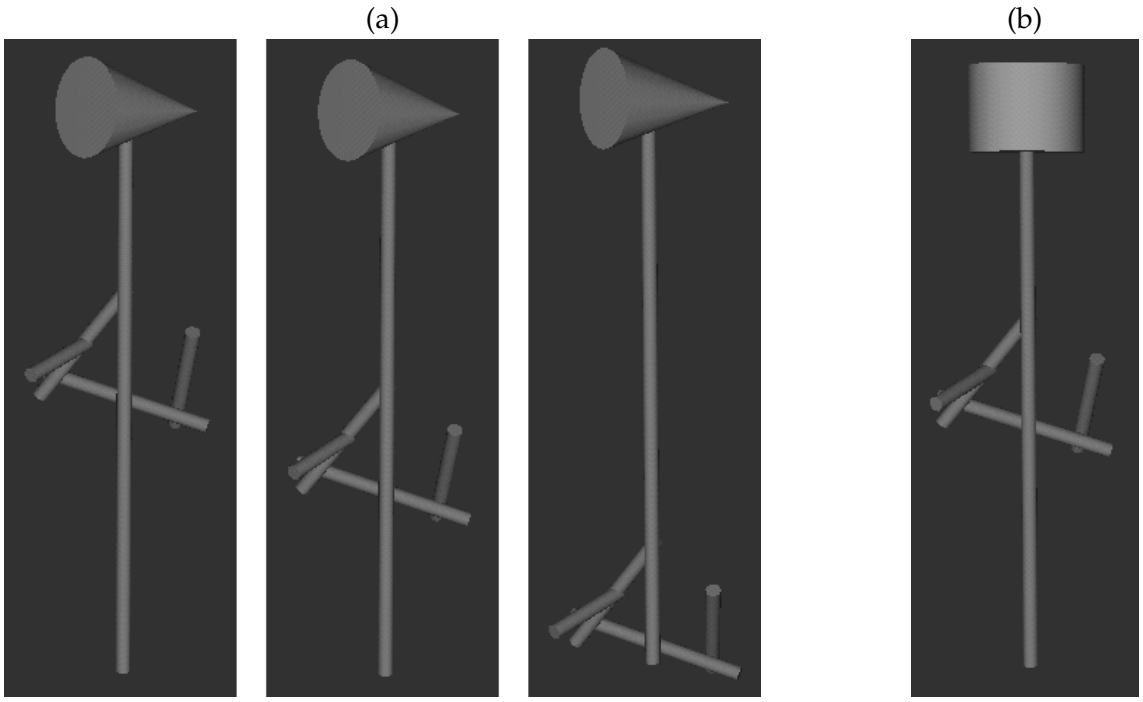

Figure 6: Shown are (a) the cone object of Figure 5, with the handle sliding along the vertical pole; and (b) the control object, with a cylinder replacing the cone.

the blocks alternated between the two object types, and the object type shown first was balanced across observers. In pilot studies we found that reaction times decreased markedly between the first and second pairs of blocks. Reaction times continued to decrease over the remaining blocks, but this decline was gradual. Thus the first pair of blocks was treated as practice, leaving a total of 120 trials per condition.

3.3 Results and Discussion

The results for the discrete-views version of the experiment are shown in Figure 8(a). On the vertical axis of the left graph is the time required to judge the color of the band on the object's handle. Recall that the appearance of the handle changed markedly across trials as the object rotated in depth. Because the cone's appearance changed in a correlated way, this part could serve as a cue to the handle's appearance. In contrast, because the cylinder did not change in appearance, it could not serve as a cue. Thus if observers could use the appearance of the cone to predict the appearance of the camouflaged handle, and if this prediction could facilitate search, then response times for the cone (white bars) should be faster than those for the cylinder (black bars). The reaction time data support this expectation (paired t-test: $t = 4.75$, $p = 0.018$). At the same time, the accuracy data do not vary systematically across the two conditions, Figure 8(b). These results indicate that appearance cuing can occur when observers are trained and tested on a few discrete views of the object.

The results for the continuous-views version of the experiment are shown Figure 9(a). Again

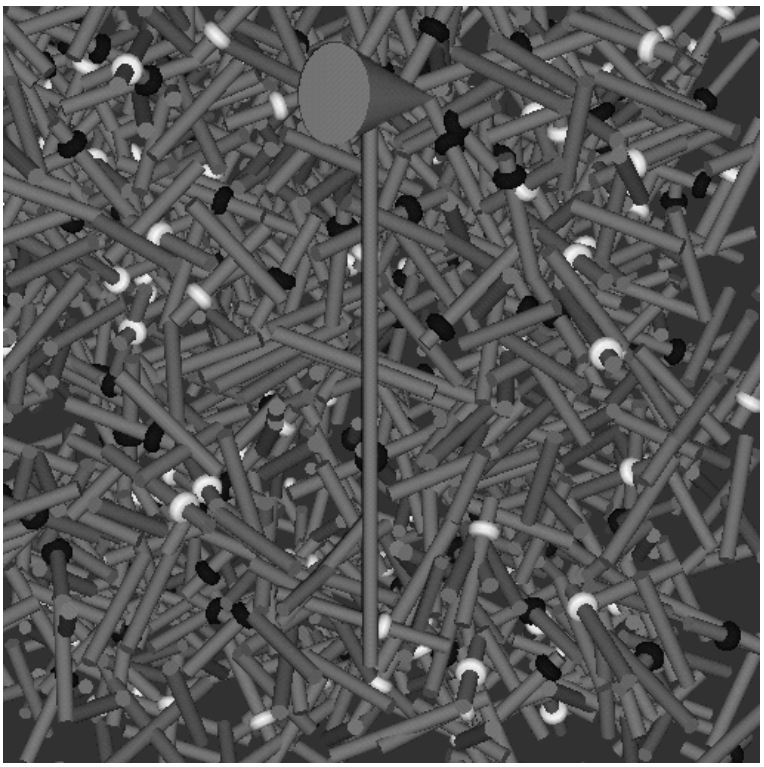

Figure 7: The cone object of Figure 5 presented against a background that camouflaged the handle. The observers' task was to find the handle and report whether it had a black or white ring.

the vertical axis reflects the time required to find the camouflaged part of the object. As before, we would expect that if observers can use the appearance of the cone to predict the appearance of the handle, then response times for the cone object should be faster than those for the cylinder object. These data do not support this expectation: there was not a significant difference in the response times for these two conditions. Figure 9(b) shows the response accuracy for the four observers.

A comparison of Figures 8 and 9 shows that the two versions of the experiment produced different results. Figure 8 indicates that when observers were trained and tested on a few discrete views of the object, the appearance of the cone facilitated search for the handle. But Figure 9 indicates that when observers were trained and tested on a continuum of views, the appearance of the cone did not facilitate search for the handle.

These results may be reconciled by considering the task demands for the two cases. In both cases the observer must identify the appearance of the cone and associate it with a particular appearance of the handle. In the discrete-views version of the experiment, it seemed unlikely that observers would confuse the three very distinct appearances of the cone or the three very distinct appearances of the handle. It also seems unlikely that they would forget the association between the cone and the handle, especially if the observers perceived the three stimuli as three distinct patterns rather than as three views of the same object. In contrast, the task demands for the continuous-views experiment were much greater. To use the cone as a cue for the handle in this case, an observer would need to make a very fine discrimination of the cone's appearance

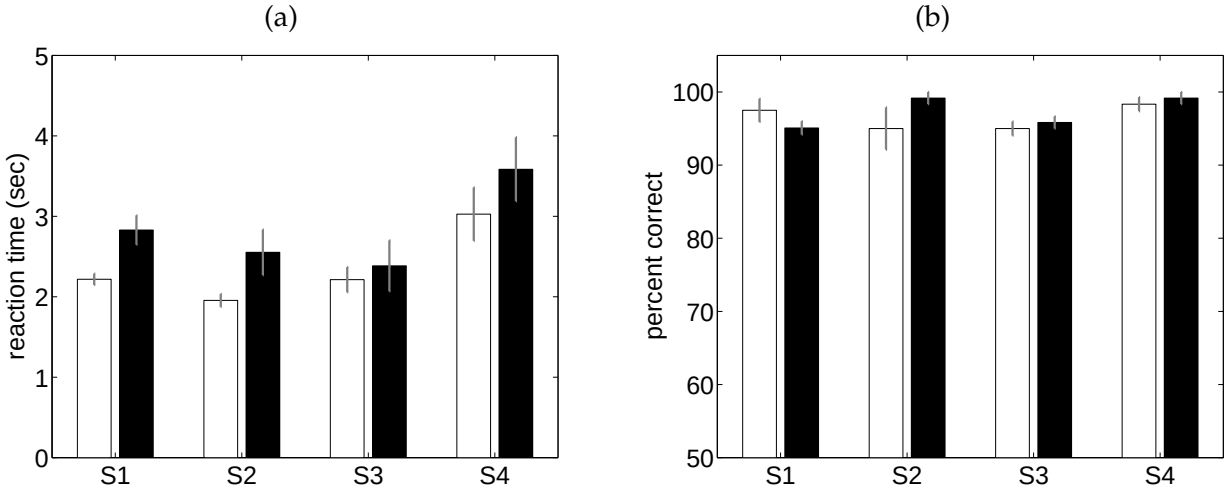

Figure 8: Results from Experiment 2 (discrete views). Shown are (a) the reaction times and (b) the response accuracy of four observers. The white bars correspond to the cone object, the black bars to the cylinder object.

and then associate it with one of many similar views of the handle. Both steps are likely to be subject to error. Thus it is very likely that the accuracy of the observers' predictions in the first discrete-views experiment was greater than that for the continuous-views experiment. To obtain an effect in the continuous-views version, it may be necessary to explicitly train observers to match the appearance of the cone with the appearance of the handle and hone their performance with feedback.

4 Experiment III: Rotations in Depth, Again

In a final experiment, we repeated the continuous views version of the experiment, but we gave observers explicit training on the association between the appearance of the cone and the appearance of the handle. To establish a criterion for this training, however, we first needed some measure of the relationship between cue accuracy and cue effectiveness. We obtained this measure by running a priming experiment in which a salient target served as the cue for the camouflaged target. That is, the cone was replaced with a white handle which was clearly visible against the gray camouflage. On most trials, the salient cue and the camouflaged target had the same pose, but on a small percentage of the trials they were rotated relative to one another. This allowed us to measure how the effectiveness of the cue was related to the accuracy of its predictions. If cues that were inaccurate by, say, 12 degrees were ineffective, then we would set our training criterion to be less than 12 degrees. We must alert the reader that the processes involved in bottom-up cuing are thought to differ from those involved in top-down, cuing (Maljkovic & Nakayama, 2000) and that these processes may have different thresholds. Nonetheless, this bottom-up cuing experiment provides a reasonable estimate of the relationship between cue accuracy and cue effectiveness.

Clearly, for the cue to be effective, observers must be well-trained to associate the cue and target. Unfortunately, this same training may also cause the cue to become ineffective, because

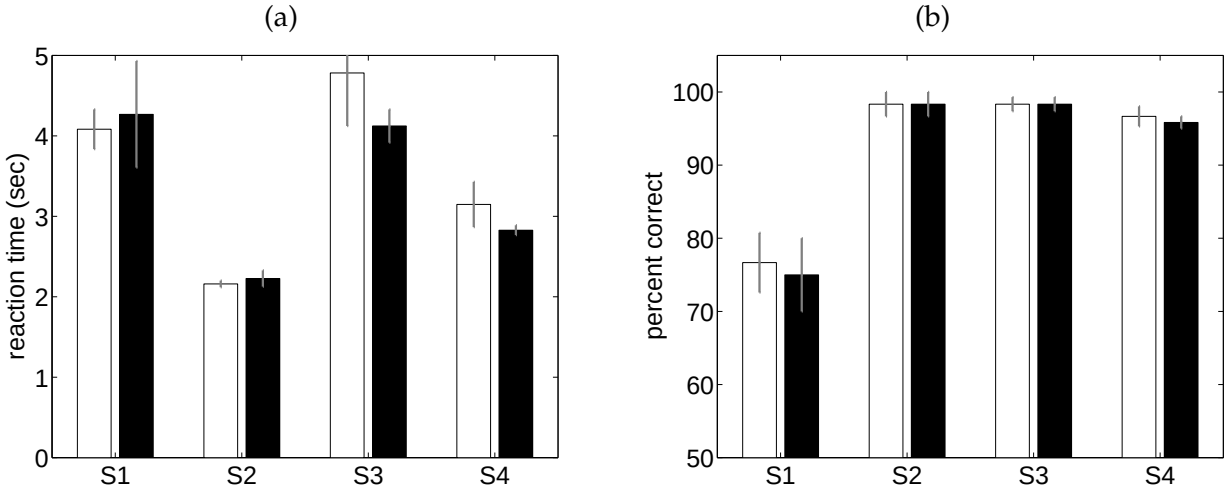

Figure 9: Results from Experiment 2 (continuous views). Shown are (a) the reaction times and (b) the response accuracy of four subjects. The white bars correspond to the cone object, the black bars to the cylinder object.

observers will simultaneously become well-trained on the target itself. Several experiments have shown that, given sufficient training on a target set, observers can search as quickly for an unspecified member of the set (e.g., any digit) as they can for a specified member (e.g., a "9") (Neisser, Novick, & Lazar, 1963; Sperling, Budiansky, Spivak, & Johnson, 1971). This kind of cost-free parallel processing is thought to occur only with suprathreshold stimuli. Signal detection theory predicts that target uncertainty should impair detection near threshold (Sperling & Dosher, 1986)and a number of experiments have borne this out (Ball & Sekuler, 1980; Greenhouse & Cohn, 1978; Davis, Kramer, & Graham, 1983). To address the possibility that observers are searching in parallel for the different appearances of the target, we added noise to our search displays. By significantly degrading the image of the target, we expected to bring this target close to detection threshold and thereby increase the effectiveness of the cue. This noise was used in the final experiment and in the preliminary experiment described next.

4.1 Preliminary Experiment

4.1.1 Stimulus

We created a new set of objects in which the cone was replaced with a high-contrast copy of the handle. Thus the object consisted of a salient handle attached by a long pole to a camouflaged handle. In some versions of this object, the pose of the top handle differed from the bottom handle by 6, 12, 18, 24 or 30 degrees, Figure 10. In still another version of the object, the cue handle was missing altogether. For this last version, the top of the pole was painted white so that it would be clearly visible against the background. (This pole indicated the horizontal location of the target but not its appearance.) A black or white ring was placed on each target handle in a random location. After these objects were added to the camouflage background, noise was added to the stimulus. The noise pattern was randomly generated from a zero-mean uniform distribution. The range of

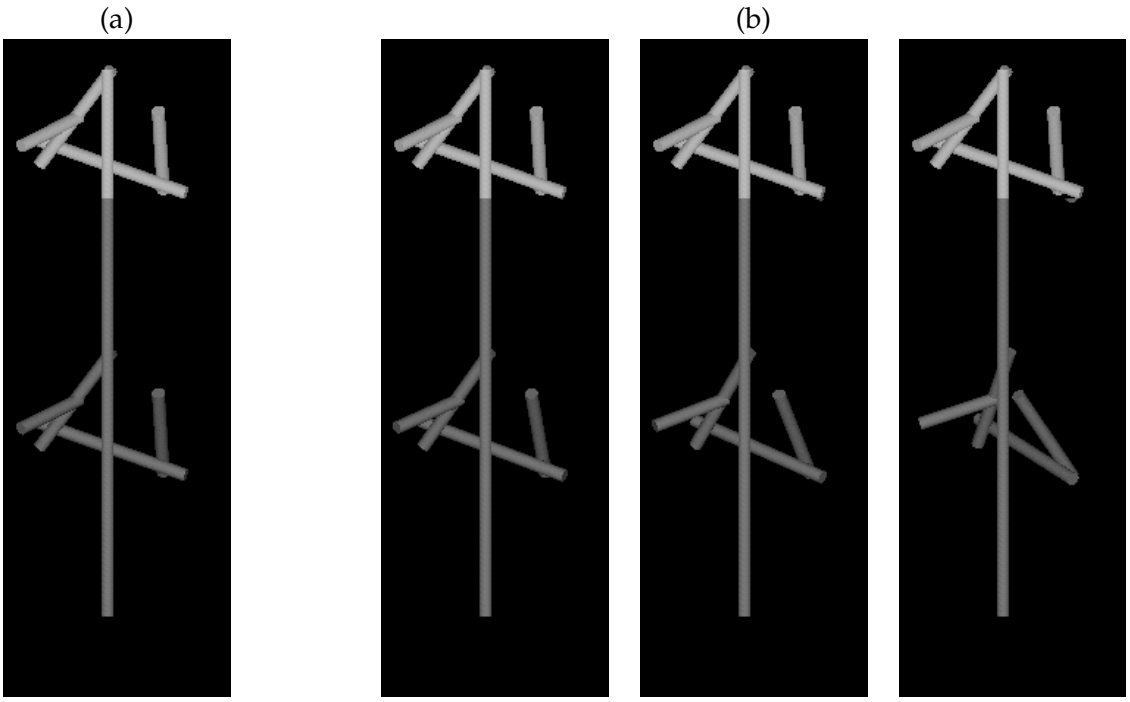

Figure 10: Examples of the objects used in the priming experiment. In (a) the two handles have the same pose, in (b) the pose of the bottom differs by 6, 12, and 30 degrees respectively. Here these objects are presented against a black background to make them easy to see, but in the experiment, they were presented with camouflage and white noise.

this distribution was gradually increased during training. When a noise pattern with a large range was added to the stimulus, some pixels occasionally exceeded allowable values ($[0, 255]$). These saturated pixels were "folded" back into the allowable range by setting pixels with gray values $p < 0$ to $-p$, and those with gray values $p > 255$ to $510 - p$. In order not to obscure the white cue or the black and white rings, this noise pattern was added only to those pixels with gray value greater than 20 and less than 235. Our monitors were not gamma-corrected and so we cannot report the signal to noise ratio of these stimuli.

4.1.2 Procedure

Six observers ran three one-hour sessions. During each session, their task was to find the camouflaged handle and to report whether it had a black or white ring. As before, the top part of the stimulus, which contained the cue, was displayed for 500 msec before the presentation of the bottom part of the stimulus, which contained the target. During their first one-hour session, the cue was always a high contrast version of the target (i.e., the cue and target had the same pose). On the first block of 18 trails, no noise was added to the stimuli. On subsequent blocks we varied the noise level to find the observer's 80% threshold. This noise level was used on the second and third days of the experiment. On these days, the observer ran a total of 22 blocks of 27 trials, 18 of these trials had accurate cues, 5 had inaccurate cues, and 4 had no cue.

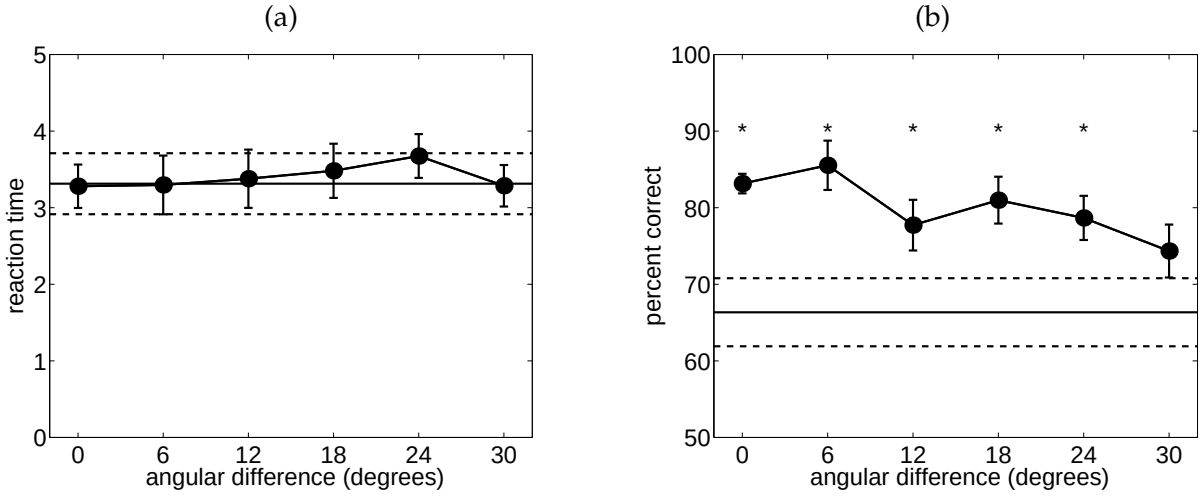

Figure 11: Results from the priming experiment. Shown are the average reaction times (a) and response accuracies (b) for five observers. The horizontal axis indicates the angular difference between the pose of the cue and the pose of the target. The horizontal line corresponds to the no-cue condition and the dashed lines reflect one standard error above and below the average. The asterisks indicate significant differences.

4.1.3 Results

Shown in Figure 11 are the average response times and accuracy rates for five observers. (The data of the sixth observer showed a speed-accuracy trade-off and so was not included.) There are two things to note from these data. First, the cuing condition had a significant effect on accuracy levels but not on response times. This increase in errors without a concomitant increase in response time may reflect misidentifications. That is, observers may have been mistaking some part of the background camouflage for the target. In any case, in our final experiment, we used accuracy as our measure of performance. The second thing to note in these data is that cues that differed from the target by less than 24 degrees provided some benefit to performance. Thus our training criterion for the final experiment was the ability to match the poses of the cone and handle to within 24 degrees.

4.2 Training

For this final experiment, six new observers were trained to match the appearance of the cone with the appearance of the handle. This training involved four one-hour sessions over a two week period. Observers began by using the arrow keys of the computer to rotate the object around its vertical axis. After 10 minutes of this free-viewing, they then began a series of matching exercises. In these exercises, the observers were presented with two versions of the object: one correct and one twisted. The pose of the cone in the two versions was the same, but the handle in the twisted version was rotated relative to the cone. The observer's task was to determine which of the two objects was correct. Auditory feedback was given throughout the training sessions. During their first session, observers were allowed to use, as a guide, the freely rotatable object.

During the remaining sessions, they were only allowed to use this guide on their first block of trials, on subsequent trials they were required to respond from memory. Over the course of the first two sessions, the matching exercise became progressively more demanding as the amount of twist was reduced from 36 degrees to 18 degrees in 6 degree steps. Observers were required to reach an accuracy level of 80 percent correct on each block of trials before preceding to the next block. Two observers needed to repeat the 18 degree block on the second training day, but otherwise all observers proceeded easily through the training exercises. On the third and fourth days of training, observers ran mixed blocks in which the decoy object was twisted by 6, 12, 18, 24 or 30 degrees. The psychometric functions shown in Figure 12(a) were collected during the fourth and final session.

4.3 Testing

After completing the training sessions, the observers returned for the testing session. The testing session was similar to that of the previous continuous-views experiment except that white noise was added to the display and the duration of the stimulus was fixed at 4 seconds. Thus, in this final experiment we measured the effect of the cue on the detectability of a degraded target image. The noise level was set at the average noise level used in the preliminary experiment. The cone and cylinder objects were painted white so that they would not be obscured by the noise. We gave observers two practice blocks, one without noise and one with noise. And because this detection task seemed quite arduous, we ran shorter blocks (18 rather than 30 trials per block).

4.4 Results and Discussion

During training, observers learned to discriminate correct objects from twisted objects in which the cone and handle had different poses. Figure 12(a) shows the average discrimination performance for observer plotted against the rotational angle between the cone and handle. All observers were performing with high accuracy ($> 85\%$) for rotation differences of 18 degrees or greater. In our preliminary study we found effective priming with cues that differed from the target by 24 degrees or less. Thus all six of these observers met our training criterion.

The data for the testing phase are shown in Figure 12(b). As in our previous version of this experiment, we did not find a significant difference between the cone and cylinder objects, suggesting that observers were not using the appearance of the cone as a cue for the appearance of the handle.

5 General Discussion

We started with the assumption that it is not always possible to segment entire objects through bottom-up processes (Barrow & Tenenbaum, 1981; Marr, 1982; Ullman, 1997). In such cases, observers may adopt a hypothesize-verify approach to object recognition. According to this approach, observers may use a distinctive part in the image to make a hypothesis about the identity of an object in the scene (Roberts, 1966; Lowe, 1985; Selinger & Nelson, 1999). This hypothesis would allow the observer to predict other object parts that should appear in the image. If the parts are found to exist, then the hypothesis is verified and the object is recognized.

Our experiments do not directly test the validity of this hypothesize-verify approach. Because our observers knew the identity of the object on each trial, we effectively by-passed the first (and

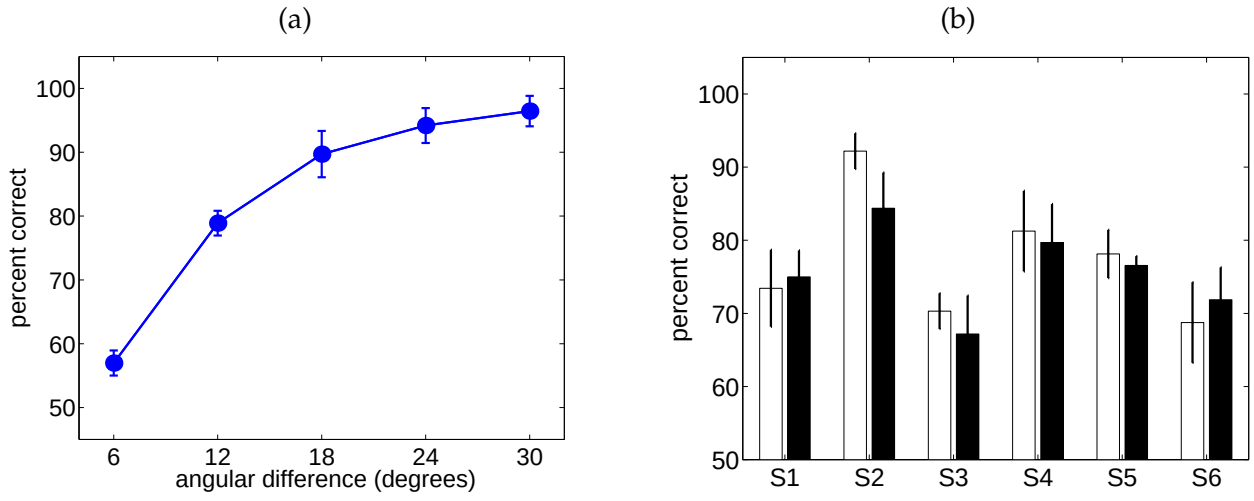

Figure 12: (a) Training results: discrimination performance as a function of the angular difference between the pose of the cone (the cue) and the handle (the target). (b) Testing results: detection performance for the cone object (white) and the cylinder object (black). Both objects were presented with both camouflage and noise.

clearly more difficult) hypothesize step. Instead, the experiments were designed to explore the nature of the predictions observers make in the verification step. We reasoned that the precision of these predictions should depend on, among other things, the symmetry of the distinctive part. Distinctive parts that have no rotational symmetry can specify the pose of the object. Thus these parts can serve as reliable cues to the location and appearance of other object parts. At the other extreme are distinctive parts that have infinite rotational symmetry. Such parts cannot specify the pose of the object and so cannot serve as such a cue.

We tested this idea in two experiments. In both, observers learned several multi-part objects by viewing them against a blank background. The observers then viewed the objects against a cluttered background that caused some object parts to be camouflaged while leaving other parts salient. The pose of the object was randomly varied across trials. Across blocks of trails, we varied the salient part's symmetry with respect to the axis of object rotation and we measured how long it took observers to find the camouflaged part. We expected that when the salient part had a low order of rotational symmetry, response times would be faster than when the salient part had a high order of rotational symmetry.

The first experiment involved rotations in the image plane. Here we found that response times were correlated with the rotational symmetry of the salient object part when the order symmetry was low: response times increased as the rotational symmetry increased from 1 to 2 to 4. But when the rotational symmetry was increased to infinity (i.e., the salient part was a circle), response times decreased. We think this non-monotonicity occurs because prediction testing is a relatively efficient search strategy when the number of predictions is low. But as the number of predictions increases, testing these predictions becomes less efficient than searching globally for the target.

In this first experiment, in which the objects were rotated in the image plane, we think it is very likely that the observers were using the appearance of the salient object part to predict the

location of the handle. That is, the salient part was functioning as a kind of spatial cue (Posner, 1980; Jonides, 1980; Cheal & Lyon, 1991). It is less clear whether this part also functioned as an appearance cue. In the second experiment, the objects rotated in depth, and while this rotation had a minimal effect on the location of the handle it had a dramatic effect on its appearance. We again found that when the salient object part had zero rotational symmetry, search times for the handle were faster than when the salient part had infinite rotational symmetry. But the expected performance benefits occured only when observers were trained and tested on a few stimulus views. When the observers were tested on a continuum of views, the effect of appearance cuing was eliminated. We repeated the continuous-views version of the experiment a second time with different training regimen and a different performance measure, but we still failed to find an effect of appearance cuing.

The most obvious difference between the discrete-views and continuous-views experiment is a quantitative one: in the first case observers were trained and tested on three very different views, in the second, they were trained and tested on 31 similar views. Clearly, the second case is much more demanding. But even after we repeated the continuous-views experiment with a more rigorous training regime, we still did not find an effect of appearance cuing. [6] There is also a second, qualitative difference between the two conditions. In the discrete-views experiment, observers may have treated the three stimuli as three distinct objects. In this case, rather than encoding three views of a 3D object, observers may have simply encoded three, 2D patterns. In the continuous-views experiment, observers clearly encoded the stimuli as 31 views of a single 3D object. Thus, the existence of appearance cuing may depend on whether the stimuli are encoded as separate objects or as different views of the same object.

Our failure to find appearance cuing would seem to conflict with a recent study showing that viewpoint cues can facilitate object recognition from novel views (Christou, Tjan, & Bulthoff, 2003). In this experiment, observers studied a set of objects from a few viewpoints. The observers then judged whether a test object presented from a non-studied viewpoint was a member of the training set. Observers were more accurate when they were first given information about the new viewpoint. While the Christou experiment and our experiment both examine whether viewpoint information can facilitate task performance, the tasks themselves are quite different. Their subjects were asked to determine if a stimulus they had never seen before was a studied object presented from a novel view. It is conceivable that the viewpoint cue might benefit performance if it allowed observers to mentally rotate even just a portion of the object to a pose similar to that of the cue. In our case, subjects were asked to find a familiar view of a familiar object in a highly cluttered scene. For the viewpoint cue to benefit performance in our case, it might be necessary for observers to use the cue to recall the appearance of the whole object with high precision. Since we do not know how recognition occurs in either case, these comparisons are purely speculative. It is clear, however, that the two recognition tasks make very different demands on the subjects.

To return then to our original question, what do our results suggest about the hypothesize-verify approach to recognition? Our results suggest that in the verification step, observers predict the location, but not the appearance, of object parts. This might indicate that our internal object representations have less viewpoint consistency than is often assumed. Observers may encode an object's features without precisely encoding the relative 3D orientations of these features. [7]

[6] It is also conceivable that we overtrained the observers, and that this made cuing less effective. As discussed earlier, we attempted to address this concern by adding stimulus noise in our final version of the experiment.

[7] One might argue that our matching results from experiment 3 showed that observers did encode the relative 3D orientations of the object parts. But it is important to note that this task involves recognition (i.e., selecting between two

Thus instead of encoding objects as rotatable 3D models (Biederman, 1987) or as a collection of 2D templates (Poggio & Edelman, 1990), it is possible that observers encode objects as a set of features with a rough spatial arrangement (Burl, Weber, & Perona, 1998). The features may be encoded with precision, but their relative 3D poses are not.

Acknowledgments

This work was supported by NSF grant SBR-9729015 (Bravo), and an Alfred P. Sloan Fellowship, NSF CAREER award IIS-99-83806, and a departmental NSF grant EIA-98-02068 (Farid).

6 References

Ball, K., & Sekuler, R. (1980). Models of stimulus uncertainty in motion detection. *Psychological Review*, *87*, 435–469.

Barrow, H., & Tenenbaum, J. (1981). Computational vision. *Proceedings of the IEEE*, *69*, 572–595.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.

Borenstein, E., & Ullman, S. (2002). Class-specific, top-down segmentation. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *European Conference on Computer Vision (ECCV)*. Springer-Verlag.

Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. *Journal of Vision*, *3*, 413–422.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Bravo, M., & Farid, H. (2003). Object segmentation by top-down processes. *Visual Cognition*, *10*, 471–491.

Burl, M., Weber, M., & Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. *European Conference on Computer Vision (ECCV)* (pp. 628–641).

Cheal, M., & Lyon, R. (1991). Central and peripheral precuing of forced-choice discrimination. *Quarterly Journal of Experimental Psychology*, *43A*, 859–880.

Christou, C. G., Tjan, B. S., & Bulthoff, H. H. (2003). Extrinsic cues aid shape recognition from novel viewpoints. *Journal of Vision*, *3*, 183–198.

Davis, E., Kramer, P., & Graham, N. (1983). Uncertainty about spatial frequency, spatial position or contrast of visual patterns. *Perception and Psychophysics*, *32*, 20–28.

Greenhouse, D., & Cohn, T. (1978). Effects of chromatic uncertainty on detectability of a visual stimulus. *Journal of the Optical Society of America*, *68*, 266–267.

Jonides, J. (1980). Towards a model of the mind's eyes movements. *Canadian Journal of Psychology*, *34*, 103–112.

Lowe, D. G. (1985). *Perceptual organization and visual recognition*. Kluwer Academic.

alternatives) rather than recall (e.g., mental imagery).

Maljkovic, V., & Nakayama, K. (2000). Priming of pop-out: III. a short-term implicit memory system beneficial for rapid target selection. *Visual Cognition*, *7*, 571–595.

Marr, D. (1982). *Vision*. W. H. Freeman and Company.

Neisser, U., Novick, R., & Lazar, R. (1963). Searching for ten targets simultaneously. *Perceptual and Motor Skills*, *17*, 955–961.

Nelson, R. C., & Selinger, A. (1998). A cubist approach to object recognition. *In Proc. International Conference on Computer Vision*, *ICCV98*, 614–621.

Pelli, D. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

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

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.

Roberts, L. (1966). Machine perception of three-dimensional objects. In J.Tippet, et al. (Ed.), *Optical and Electro-optical Information Processing*. MIT Press.

Selinger, A., & Nelson, R. C. (1999). A perceptual grouping heirarchy for appearance-based 3D object recognition. *Computer Vision and Image Understanding*, *76*, 83–92.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*, 29–56.

Sperling, G., Budiansky, J., Spivak, J., & Johnson, M. (1971). Extremely rapid visual search: The maximum rate of scanning letters for the presence of a numeral. *Science*, *174*, 307–311.

Sperling, G., & Dosher, B. A. (1986). *Handbook of human perception and performance: Volume 1*, Chap. Strategy and optimization in human information processing. Wiley Press.

Ullman, S. (1997). *High-level vision: Object recognition and visual cognition*. Bradford/MIT Press.

Ullman, S., Sali, E., & Vidal-Naquet, M. (2001). A fragment based approach to object representation and classification. In C. Arcelli, L. Cordella, & G. S. di Baja (Eds.), *Visual Form 2001, 4th International Workshop on Visual Form, IWVF-4, Capri, Italy*.