Visual Cognition, 10(4):471-491, 2003

Object segmentation by top-down processes

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

In cluttered scenes, some object boundaries may not be marked by image cues. In such cases, the boundaries must be defined top-down as a result of object recognition. Here we ask if observers can retain the boundaries of several recognized objects in order to segment an unfamiliar object. We generated scenes consisting of neatly stacked objects, and the objects themselves consisted of neatly stacked colored blocks. Because the blocks were stacked the same way within and across objects, there were no visual cues indicating which blocks belonged to which objects. Observers were trained to recognize several objects and we tested whether they could segment a novel object when it was surrounded by these familiar, studied objects. The observer's task was to count the number of blocks comprising the target object. We found that observers were able to accurately count the target blocks even when the target was surrounded by up to four familiar objects. These results indicate that observers can use the boundaries of recognized objects in order to accurately segment, top-down, a novel object.

Object Segmentation   Object Recognition   Visual Working Memory

[1] Author to whom correspondence should be addressed: Department of Psychology, Rutgers University, Camden NJ 08102. tel:856.225.6732; 856.225.6602; email: mbravo@camden.rutgers.edu

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

1 Introduction

We perceive a world organized into familiar objects such as can openers, tractors and picnic tables. Objects are defined by the connectedness of their parts, and we learn about this connectedness through experience [14, 23, 1]. We perceive a can opener as an object because we have noticed that when we pick up one of its parts, all of the other parts come with it. We perceive a tractor as an object because we have noticed that when one part of the tractor moves, all of the other parts move with it. And, although we may have never seen it move as a whole, we perceive a picnic table as an object because we have noticed that its parts always appear together.

While we learn about the connectedness of object parts from visual experience, this connectedness is not always easy to discern in a given visual image [2, 11, 26]. For example, when we look into a kitchen drawer, the parts of the can opener may be contiguous not only with each other but also with parts of the potato masher and ice cream scoop. Admittedly, there may be other visual cues that help us segment these cooking utensils: the ice cream scoop may be white, the potato masher black and the can opener silver. But these similarity cues are not always present, and they can even be misleading when an object is made from several materials.

The line drawing in Figure 1 illustrates this problem. We perceive this drawing as representing two objects, a lamp and a table, even though there are no clear visual cues that cause us to organize the drawing this way. The ambiguity depicted in this line drawing arises frequently in the cluttered scenes we view every day. That is, when objects are stacked on top of or next to one another, there may be no image cues differentiating parts that are truly connected from parts that are merely adjacent. And when an observer is permitted only a single, monocular view of a scene, there may be no image cues differentiating parts that are adjacent in the world from parts that are only adjacent in the image.

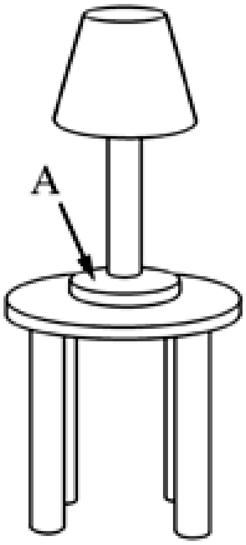

Figure 1: Why does the part labeled ’A’ group with the cylinder above it and not the one below?

This ”part ownership” ambiguity that arises in cluttered scenes clearly poses a problem for models of human object recognition because these models presuppose a segmented object. A few computer vision models have attempted to deal with the problems posed by clutter. One such approach, termed hypothesize-verify, involves the following steps (e.g., [21, 9, 5, 22]). Bottom-up segmentation processes (e.g., the Gestalt grouping rules) first delineate features in the image. Depending on the model, the complexity of these features ranges from contours to surfaces to volumetric parts. A subset of these features is then matched against object representations in memory. The matches are used to predict other features of the object that should appear in the scene. If the predictions are correct and the existence of these features is verified, the object is recognized.

We start then with the assumption that complete object segmentation [3] is not a necessary prerequisite for object recognition. We also assume that object segmentation is not a necessary consequence of object recognition. On the surface, this second assumption might seem odd. It might seem more natural to suppose that once you have recognized an object you would automatically segment it from its background. But consider that recognition need not be based on the whole object, some distinctive feature or part may suffice. (Imagine coming across an alligator in the grass; once you recognize a patch of scaly skin or a beady eye, you may not find it necessary to fully segment the entire animal before reacting.) Furthermore, if recognition involves a purely feed-forward process, then precise information about the locations of object parts may be discarded in this process.

The goal of the experiments reported here was to examine one of the ways in which object recognition can lead to object segmentation. Once an object has been recognized, we ask whether observers can mark its parts or boundaries for exclusion when segmenting neighboring objects. Essentially, we want to know whether observers can "subtract-out" familiar objects in order to find the boundaries of an unfamiliar object.

2 Experiment 1

To test whether object recognition can drive object segmentation, we developed stimuli with ambiguous part ownership. The stimuli we used consisted of realistically rendered blocks which were neatly stacked to form objects, Figure 2. Our observers studied the individual block-objects during several training sessions. They were then presented with scenes composed of a number of these block-objects stacked next to one another, Figure 4. Each scene also contained a novel block-target positioned between two of the familiar objects, Figure 3. The observer's task was to find this target and count its blocks. Because the objects and target were all neatly stacked next to one another, there were no image-based cues indicating which blocks belonged to the target and which to the objects. Thus to perform this task, observers were forced to rely on their knowledge of the familiar objects to locate and count the target blocks.

Since we asked observers to find the novel target amongst familiar objects, it might appear that an observer could have simply searched for something unfamiliar. When two objects were placed next to one another, however, new block configurations were created at the boundary between them. And because the observer's task was to report the number of target blocks, accurate performance required the observer to find the boundaries of the familiar objects.

This first experiment simply examines whether observers can perform the task. If they can, this would show that observers can recognize objects that they cannot fully segment from image cues. More importantly, it would show that they can keep track of the boundaries of recognized objects in order to segment an unfamiliar object.

2.1 Methods: stimuli

The stimuli were created using OpenGL on an SGI O2. To optimize the three dimensional appearance of the displays, the stimuli were rendered under a perspective projection with directional and ambient lighting. A dark gray plane defined the ground, the remaining background was light gray. The stimulus rotated at 30 deg/sec about a central vertical axis. This rotation provided strong motion parallax cues for scene structure and it permitted viewing from multiple vantage points. Except for the free-viewing condition (see Training below),

[3] Note that we define object segmentation as the segmentation of the entire object (the entire can opener, the entire lamp) from its background.

all stimuli were viewed from an elevation of 40 degrees.

2.1.1 Objects

Block objects were constructed by a fully automated computer program which neatly stacked seven blocks. The program first placed one block on the ground plane and then randomly selected one of the five visible facets for the placement of the second block. The program then selected one of the visible facets for the placement of the third block. This process was repeated four more times as additional blocks were added to random facets of the growing object. The program imposed two restrictions on the placement of the blocks. First, before putting a block along a side facet, the program checked that the block would be supported by another block or by the ground plane. We wanted the connectivity of all the blocks to be ambiguous, and we assumed that an unsupported block would appear to be physically attached to the block at its side. Secondly, the program required that each object have at least one block stacked on top of another block. Thus, any object that consisted entirely of blocks on the ground plane was automatically rejected. This requirement was based on our subjective impression that flat objects looked less object-like than objects with more complex 3D shapes.

The seven blocks in each object were randomly assigned one of four highly saturated colors (red, blue, green or yellow). The blocks were separated by a small gap so that neighboring blocks with the same color would not appear to merge into a single elongated block. However, the gaps were so small that they did not prevent the blocks from appearing connected.

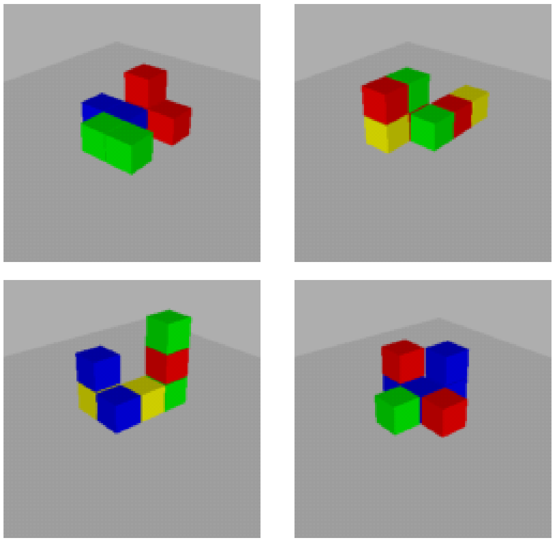

Figure 2: Four objects each constructed by neatly stacking seven colored blocks. Some blocks may be difficult to see in these black and white images, however, all blocks were clearly visible in our color stimuli.

Twelve objects were selected from the output of the program and used for training (Figure 2 shows four of these objects.) Two subjective criteria were used in selecting the training objects. First, we wanted them to be fairly easy to discriminate, and so we chose a range of shapes and color patterns. Second, we wanted them to look like average objects. Thus, we did not select objects with highly distinctive features such as tall vertical stacks or large clusters of similarly colored blocks. These hand-selected objects were divided into two sets of six objects. One set was used for the training of observers S1 and S2 and as the unfamiliar control for observers S3 and S4. The other set was used for training observers S3 and S4 and as the unfamiliar control for S1 and S2.

2.1.2 Targets

Targets were generated in exactly the same way as the objects, but only four, five or six blocks were used. A new target was generated for each trial. Shown in Figure 3 are example targets. We did not check that each of these randomly generated targets was unique, but given the many stacking and coloring permutations it is likely that most were.

Because we wanted to compare performance for top-down segmentation with bottom-up segmentation, we also created targets that were defined by an image cue. The blocks comprising these targets had a lower reflectance than the object blocks. That is, because these target blocks were assigned lower RGB levels than the object blocks, they appeared darker. The RGB level for the object blocks was 250 for all colors (250,0,0 for red, 0,250,0 for green, 250,250,0 for yellow and 0,0, 250 for blue). The RGB level for the target blocks was reduced by 10, 20, 30, 40, 50, or 100. We determined in a pilot experiment that with this span of RGB levels, target discriminability ranged from sub-threshold to clearly supra-threshold.

We should note that the discriminability of the object and target blocks was related to RGB level in an complex way. This is because discriminability probably depended not only on RGB value but also on the block's color and the block's context (i.e., whether it was next to a similarly colored block). And since the RGB level determines reflectance, the luminance of each facet also varied with the illumination (i.e., with the angle between the surface normal and the directional lighting source). Despite these potential sources of variability, the overall relationship between RGB value and discriminability was monotonic. Thus, by varying RGB level we were able to systematically vary the strength of bottom-up segmentation.

2.1.3 Scenes

All scenes contained a target and four objects. A different scene was created for each trial by a fully automated program which implemented the following steps. First, the target was positioned on the ground plane and two of its sides (north, south, east or west) were randomly selected. An object was then positioned along each of the selected sides. This ensured that observers would have to identify at least two object boundaries in order to locate the target blocks. A bounding box was drawn around this scene, and facets touching the box were noted. One of these bounding facets was randomly selected and a third object was placed next to it. This process was repeated again for the placement of the fourth object. Before being added to the scene, each object was randomly rotated about its vertical axis by 0, 90, 180 or 270 degrees. The four objects in each scene were selected randomly but without replacement from the training set so that no object appeared more than once in a scene. In the final step of the program, the scene's center of mass was moved to the middle of the ground plane. It is important to note that the target and objects were neatly stacked in the scene so that adjacent blocks were aligned the same way within and across the objects and the target, Figure 4.

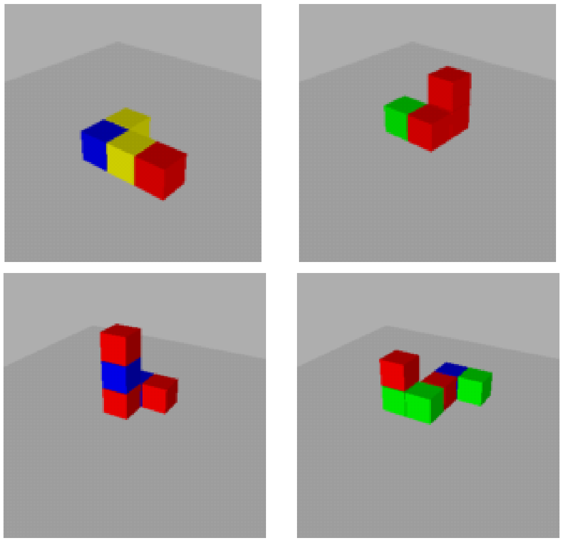

Figure 3: Four example targets constructed by neatly stacking four, five or six colored blocks.

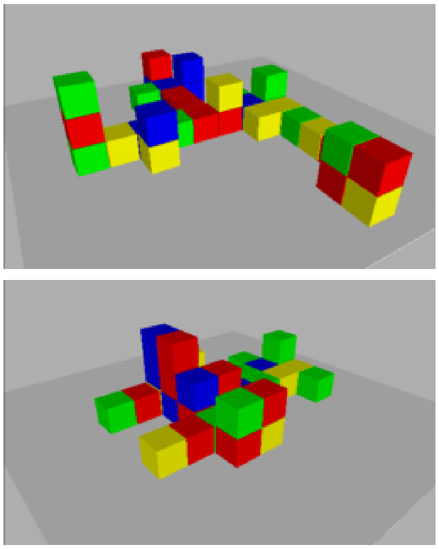

Figure 4: Snapshots from two of the rotating scenes. Each scene was constructed by neatly stacking four objects and one target. Note that the boundaries within and across objects are identical.

2.2 Methods: procedure

2.2.1 Training

Observers learned six objects during four, one-hour training sessions. Each session included self-paced exercises in which observers freely viewed rendered objects or used wood blocks to build real objects. Each session also included tests of the observers' knowledge of the objects, with the tests becoming more difficult over time. The four training exercises are described below.

1. Free viewing of the objects. The observers started each session by viewing the individual objects on the computer screen. Each object was given a name (sed, tran, mats, elel, choo and halb) and a number (1, 2, 3, 4, 5 and 6) and these identifiers appeared at the top of the screen when the object was displayed. Using the arrow keys on the computer keyboard, the observers were able to freely rotate the object through any azimuth and positive elevation. The objects rested on the ground plane, and so observers never viewed them from the bottom. This part of the training was entirely self-paced: the observers could choose to view any one of the objects at any time, and the exercise was terminated when observers felt ready to try the identification task. Typically, observers studied the objects for 20-30 minutes on the first day and 5-10 minutes on subsequent days.

2. Identifying the objects. An object was displayed on the computer screen and the observer was asked to identify it by pressing the appropriate key on the keyboard. The keys were labeled with both the object's name and number. Auditory feedback was given after an incorrect response. A test consisted of 36 trials, and observers were required to pass six tests of increasing difficulty. In the first test, an object rotated through 360 degrees in 10 seconds. On the second and third tests, a viewing azimuth was randomly selected, and a static object was displayed for 1 second and 0.5 seconds, respectively. On the fourth, fifth and sixth identification tests, the object rotated though 360 degrees in 10 seconds as in the first test. But now additional blocks were stacked next to the object, and the positions and colors of these distractor blocks changed on every trial. The number of these distractor blocks increased from three to five across the tests.

3. Discriminating the objects from decoys. An object was displayed on the screen and, on 30% of the trials, the color of one block was changed. The observers' task was to determine whether the displayed object was a studied object or a decoy.

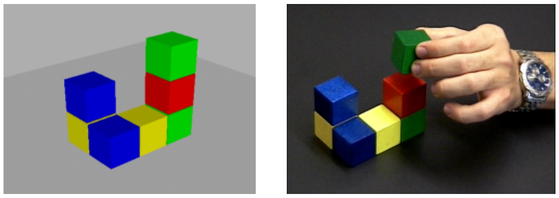

Figure 5: As part of the object learning stage observers constructed block objects from real blocks.

Again, the test was repeated six times. Just as with the identification test above, the tests were made progressively more difficult by decreasing the stimulus duration and by placing distractor blocks around the object.

4. Building the objects from real blocks. Observers built the objects from one-inch wooden blocks that were painted to match the rendered blocks, Figure 5.

In order to pass each of the identification and discrimination tests, observers were required to be correct on at least 32 of 36 trials. On the last day of training, each observer was required to quickly and accurately build the objects from memory. All four of the observers recruited for this study passed the training phase and moved onto the experiment. The experiment involved three, one-hour sessions which were completed within the same three week period as the training.

2.2.2 The experiment

In the experiment, the observer's task was to find the target in each scene and count its blocks. This task was performed under three conditions: (1) Top-down: the target was placed alongside familiar objects. Because the target blocks were indistinguishable from the object blocks, the observers had to use knowledge of the objects to segment the target. (2) Bottom-up: the target was placed alongside unfamiliar objects. The target blocks were darker than the object blocks and so the observer could use an image cue to segment the target. (3) Both: the target was placed alongside familiar objects and its blocks were darker than those of the objects. Observers could use bottom-up and top-down cues to segment the target. [4]

These three conditions were run in separate, interleaved sets of 36 trials. In all, the observer ran 27 sets, 9 for each condition, but the results for the first set of each condition were discarded as practice. The observer initiated the first trial in each set. The display appeared after a second and remained on until the observer responded by pressing the appropriate key on the keyboard (”4”, ”5” or ”6” for 4, 5, or 6 blocks in target). Feedback was provided on incorrect trials.

Observers were paid $10 per hour and were recruited from the undergraduate population at Rutgers/Camden. Three male observers and five female observers between the ages of 20 and 40 participated in the two studies reported here. None of the observers had prior experience in a psychophysical study, nor were they aware of the purpose of this study.

2.3 Results and Discussion

The data for four observers are shown in Figure 6. On the horizontal axis is the RGB level difference between the target and object blocks. On the vertical axis is percent correct (left column) or reaction time (right column). [5]

[4] We recognize that all visual tasks involve top-down and bottom-up information. Thus our top-down condition involves only relatively more top-down processing than our bottom-up condition. And the name of our Both condition refers to the fact that it is a combination of conditions 1 and 2.

[5] As is usual for reaction time data, our data had a skewed distribution. Because we did not want our results to be biased by a small number of very large reaction times, we report only reaction times that fell within

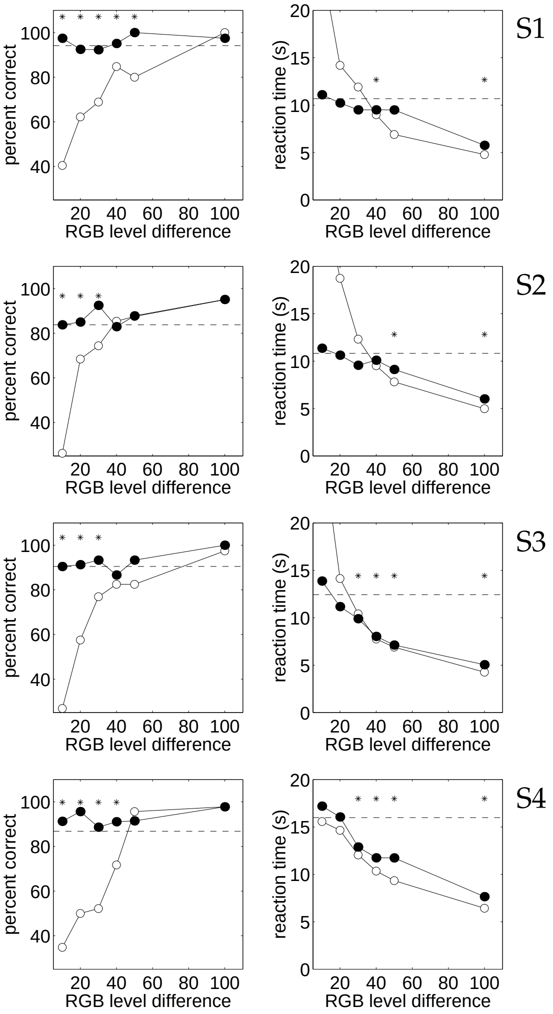

Figure 6: Speed and accuracy data for four observers from Experiment 1. The horizontal axis indicates the RGB level difference between the target and object blocks in the bottom-up and both condition. The dashed line corresponds to the top-down condition, the open circles to the bottom-up condition, and the filled-circles to the both condition. N=48 for each circle, N=288 for the dashed line. The significance of the asterisks is explained in the text.

The data for the bottom-up condition are indicated by open circles. In this control condition, scenes were composed of unfamiliar objects and a target with a lower reflectance than the objects. When the reflectance difference between the target and object blocks was very large (an RGB level difference of 100), the target was highly salient. Observers responded with a high level of accuracy in about five seconds. As the difference was reduced, accuracy fell and reaction times increased. For an RGB difference of 10, accuracy was near chance (33%). While it is not surprising that observers were unable to count the target blocks when these blocks were indistinguishable from the object blocks, it is still worth noting because it indicates that there were no inadvertent stimulus cues defining the target. (Recall that the control objects for half of the observers were the experimental objects for the other half.) This validates our claim that in the top-down condition described next, observers must have used their object knowledge to find the target.

The data for the top-down condition are represented by the dashed horizontal line in Figure 6. In this condition, the RGB difference between the target and object blocks was 0, but we have represented the data as a line so that it may be easily compared with the data from the other conditions. To find the target in these scenes, observers were forced to use a top-down segmentation strategy. For all four observers, accuracy was somewhat less than that for scenes with salient bottom-up cues, but it was still high. Responses were 2-3 times slower.

The last set of data, the filled-circles, corresponds to the "both" condition. In this condition, the objects were familiar and the target blocks had a lower reflectance value than

two standard deviations of the mean. This truncation method excluded less than 5% of the data. This exclusion rate was similar across conditions and did not affect the overall pattern of results.

the object blocks. Thus observers could use both top-down and bottom-up information to segment the target. At the extremes of the RGB range, observers used the best source of information available. When the RGB difference between the target blocks and the object blocks was small, performance resembled the top-down condition. When this RGB difference was large, performance resembled the bottom-up condition, although it was not quite as fast. [6] When the RGB difference was moderate (30 or 40), three of the four observers (S1, S2 and S4) did better in the both condition than they did in either the top-down or bottom-up conditions. That is, the both condition produced more accurate responses than the bottom-up condition and faster responses than the top-down condition. To compare accuracy levels, we computed z-values using the equation, $(P_1 - P_2)/\sqrt{pq(1/N_1 + 1/N_2)}$, where $p = (P_1 N_1 + P_2 N_2)/(N_1 + N_2)$, $q = 100 - p$, P_1, P_2 are the accuracy levels, and N_1, N_2 are the number of trials for the both and bottom-up conditions. In the graphs of accuracy, the asterisks indicate RGB levels in which the both condition was significantly more accurate than the bottom-up condition with p < 0.05. To compare reaction times, we performed an ANOVA on the means of the eight blocks of trials run by each observer. In the graphs of reaction time, the asterisks indicate RGB levels in which the both condition was significantly faster than the top-down condition (F(1,14) > 4.6, p < 0.05). We should note that this finding does not necessarily imply that observers can integrate these two very different sources of segmentation information. The task required observers to both find and segment the target, and so the top-down and bottom-up information may have contributed in different ways to the task. For example, observers may have used bottom-up information to locate the target, and top-down information to find its boundaries. [7]

The point of this experiment was simply to show that observers can use their knowledge of familiar objects to find the boundaries of an unfamiliar target. Relative to salient bottom-up segmentation, this top-down segmentation is almost as accurate, but it is much slower. This slowness may be exaggerated by our stimuli. The key design requirement for these stimuli was that they permit strict control over the apparent connectedness of the object parts. We also wanted the objects to be interchangeably stackable so that we could generate hundreds of unique scenes in an automated way. Blocks seemed like an obvious choice. But in using a single shape, we had to rely on a surface property (color) to distinguish different parts. Some have argued that surface properties play a minimal role in object recognition, while others argue that they can be important (for a review see [24]). Further, since none of the objects were associated with unique parts, they could only be distinguished by the configurations of these parts. The reliance on color patterns to distinguish these objects may have made their recognition particularly slow.

In this first experiment, the target was always located between two objects, and so observers were required to retain information about two object boundaries. It seems very likely that there is a limit to the number of object boundaries an observer can store in visual working memory. To try to find this limit we increased the number of objects surrounding the target in the next experiment.

[6] For optimal performance, observers should have ignored the familiar objects in the both condition and attended only to the bottom-up cue. Apparently, they were unable to do this. This may be because they had recently spent four training sessions responding to the directly to the objects.

[7] We also examined the effect of target size (the number of blocks in the target) on performance and found a very similar pattern for all three conditions. There was a consistent effect of target size on accuracy; observers performed better when the target contained fewer blocks. Surprisingly, target size had little effect on reaction time.

3 Experiment 2

In this second experiment, a novel target was located near the center of each scene and one to six familiar objects were placed alongside it. The observers task was again to count the number of blocks in the target. Because there were no visual cues defining the target blocks, observers had to rely on their recognition of the familiar objects to determine which blocks belonged to the target. By varying the number of objects surrounding the target, we hoped to determine whether there is a limit to the number of recognized objects that an observer can subtract-out of a scene in order to find a novel target.

3.1 Methods

Four new observers were recruited to run in this second experiment. Because the methods closely resembled those of the previous experiment, we note only the differences here. In this second experiment, scenes consisted of one to six objects positioned around a central target. In order to pack more objects around the target it was necessary to use slightly larger targets than in the previous experiment (5,6 or 7 blocks rather than 4, 5, or 6). It was also necessary to use a different method for placing the objects in the scenes. Rather than adding objects along the perimeter of a bounding box, the objects were positioned around the target and then collisions between objects were detected. If a collision did occur, the objects were automatically rotated and repositioned until a valid scene was generated.

As in the previous experiment, the target could be identified in one of two ways, depending on the condition. In the top-down condition, the target blocks were indistinguishable from the object blocks, but the observers were familiar with the objects. Thus the observers could use their object knowledge to segment the target. In the bottom-up condition, the objects were unfamiliar, but the target blocks had an RGB value of 150, while the object blocks had an RGB value of 250. Because the target was noticeably darker than the objects, observers could use a bottom-up cue to segment the target. These two conditions were interleaved in eight sets of 48 trials.

As our measure of segmentation, we asked observers to report the number of blocks composing the target. Although the counting task required that observers determine which blocks belonged to the target, it did not explicitly test whether observers perceived the target as a whole object. To test whether observers could perceive the whole target, we also ran an abbreviated version of the experiment in which observers were required to build the target, from memory, out of blocks. To reduce the observer's memory load, the number of blocks in the target was reduced to 4 or 5 (and the observers were instructed as such). The number of objects in the scene was fixed at four. Observers were allowed to inspect the scenes for as long as they wished before they started building the targets. The scene was removed from view during building and observers were not given feedback. This building task was run in two blocks of 16 trials for both the top-down and bottom-up conditions.

3.2 Results and Discussion

The data in Figure 7 show how well observers were able to count the number of blocks in the target as the number of familiar objects surrounding this target varied from one to six. We consider first the accuracy data from the bottom-up condition (open circles). In this control condition, the target blocks were noticeably darker than the object blocks, and observers could use a salient image cue to segment the target. For this condition, accuracy fell off very gradually, and was still well above chance even when six objects surrounded the target.

The filled circles correspond to the top-down condition in which observers had to rely on their knowledge of the familiar objects to segment the target. For all four observers, accuracy fell off gradually as the number of objects surrounding the target increased from one to four. When the number of objects increased to five, however, accuracy dropped off sharply, and when the number of objects was increased to six, performance fell almost to chance (33%). It appears that when observers segment a novel object by excluding familiar objects, they can reliably exclude about four objects. While it may be coincidental, it is interesting to note that other experiments using very different methodologies have shown that four objects is the upper limit of visual working memory [19, 20, 8, 10].

For completeness, we have included the reaction time data from this experiments (right column of Figure 7). As the filled circles indicate, the top-down segmentation process was very slow relative to the bottom-up segmentation process. If we exclude conditions that had an accuracy less than 75% (i.e., scenes with more than four objects in the top-down condition) then the average slope in the top-down condition was 2.3 seconds per object, while the average slope in the bottom-up condition was 0.47 seconds per object. It is important to note that these set size effects reflect both internal processing time and external (stimulus) limitations. In our three-dimensional stimuli, the targets and distant objects were sometimes partially occluded by near objects. When this happened observers had to wait for the scene rotation to bring the occluded blocks into view. The chance for such occlusions was the same for the top-down and bottom-up conditions, but it did vary with the number of objects in the scene. So while these occlusions cannot account for the reaction time difference between the top-down and bottom-up conditions, they can at least partly explain the increase in reaction times with object number for both condi-

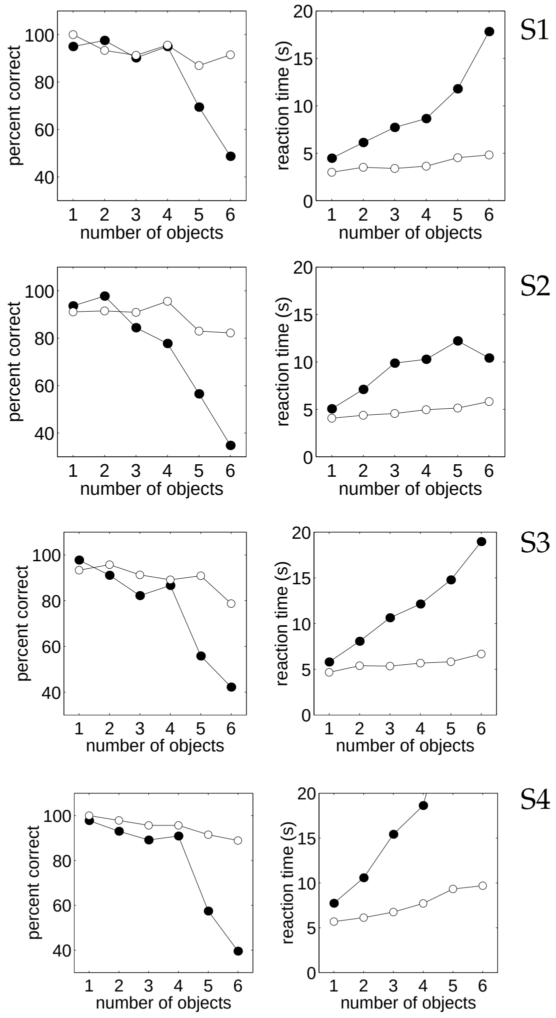

Figure 7: Speed and accuracy data for four observers from Experiment 2. The filled circles correspond to the top-down condition, the open circles to the bottom-up condition.

tions. [8]

We have assumed, that in order to accurately count the blocks belonging to the target, observers would need to isolate this unfamiliar object. But it is possible that observers may have simply moved systematically through the display, counting the blocks that did not belong to a familiar object without representing these extra blocks themselves as a distinct object. To test whether observers perceive the target as a distinct object, we also conducted an abbreviated version of the experiment in which observers were required to build the target, from memory, using colored blocks. In the displays for this building task, the target was always surrounded by four objects. The targets built by the observers were scored as correct, small error or large error. Correct responses included exact replicas of the target in the scene as well as its mirror image. Incorrect responses were divided into two kinds: small and large. Small errors involved one of the following: 1) the colors of two blocks were interchanged, 2) one block was the wrong color but this incorrect color was found elsewhere in the target, or 3) the position of a block was off by one facet. Errors were considered large if they involved two small errors or one of the following: 4) one block was the wrong color and this incorrect color was not found elsewhere in the target, 5) one block was more than one facet away from its correct position, 6) one block was missing, or 7) one block was extra.

			small	large
subject	condition	correct	error	error
S1	top-down	28	2	2
	bottom-up	28	3	1
S2	top-down	24	3	5
	bottom-up	25	4	3
S3	top-down	21	6	5
	bottom-up	25	2	5
S4	top-down	26	3	3
	bottom-up	27	3	2

Table 1: Building accuracy from Experiment 2: top-down versus bottom-up, N = 32.

As the data in Table 1 indicate, on the majority of trials observers were able to build an exact replica of the target object, from memory. Although the percentage of correct targets was slightly lower for the top-down displays (77%) than for the bottom-up displays (82%), performance was similar for the two conditions. Even when errors were made, it was always possible to determine the correspondence between the built target and the model target, and the errors were limited to those described above. Thus the results from this building task suggest that when using top-down processes to segment the novel target, observers were able to perceive the target as a whole object.

4 General Discussion

A number of researchers have noted that many objects do not exist as visual entities in the image. As Marr put it, ”regions that have semantic importance do not always have any particular visual distinction” [11] (p.270). So while bottom-up segmentation may yield contours, surfaces, or even object parts, it does not reliably yield objects. Our stimuli are extreme

[8] We have also conducted a similar experiment with stationary stimuli. In this experiment, the objects were always placed on the north, east and west sides of the target and the target was viewed from the south. Thus the target blocks were never occluded by the objects, but the total number of objects was limited to a maximum of three. With these stationary stimuli, the average slope for the top-down condition was 1.66 seconds per object and the average slope for the bottom-up condition was 0.41 seconds per object (N = 4).

examples of scenes in which the boundaries within objects are no different from the boundaries between objects. Thus the objects could not be fully segmented using bottom-up cues. Nonetheless our observers could recognize the objects in our scenes. More significantly, we found that observers could use the boundaries of several recognized objects to segment an unfamiliar object. Before we discuss this finding, we will briefly review the existing evidence showing that object recognition does not require full object segmentation.

Some of the most compelling evidence that object recognition does not require object segmentation comes from demonstrations and experiments involving quantized stimuli [7, 4, 13]. Of these, perhaps the most famous is the Dalmatian dog scene by R.C. James. This two-toned image depicts a spotted dog in a sun dappled yard. Because the scene has been quantized to two luminance values, many contours are missing from the image and there are few image cues defining the dog's boundaries. Nonetheless, observers can still recognize the dog. It should also be noted, however, that when observers are first presented with this scene, they initially perceive a field of black spots. Thus, when virtually all of the image information concerning the object boundaries has been removed, recognition is very slow.

Other demonstrations that object recognition does not require object segmentation have taken a complementary approach. Rather than eliminating object contours, the object contours are made continuous with other contours in the image [6, 12]. Thus, bottom-up segmentation processes group the contours of the object with those of the background. As with the quantized stimuli, object recognition may be greatly delayed with these embedded figure stimuli. But because observers are still able to detect the figure, this research indicates that object recognition may proceed without object segmentation.

A third line of research showing that object recognition does not require full object segmentation has focused on ambiguous figure-ground stimuli similar to Rubin's vase stimulus. In these stimuli, the object boundary is neither missing nor camouflaged, instead there is a single highly salient boundary separating a black region and a white region. But rather than perceiving this boundary as common to both regions, observers perceive one region as owning the boundary and the other region as extending behind it. The boundary-owning region is the figure, the other region is the ground. Which region owns the boundary fluctuates over time, giving rise to a bistable percept of figure and ground. Peterson has shown that the region that is most likely to be seen as the figure first and longest is determined in part by the degree to which the regions correspond to familiar objects [15, 16, 18, 17]. This research shows that border ownership, or equivalently depth-ordering, need not be resolved prior to object recognition.

Our experiments resemble these earlier experiments in that they required observers to recognize objects that they could not fully segment using only bottom-up processes. But our experiments differed from this research because we required observers to use the recognition of the familiar objects to segment an unfamiliar object. Thus, after recognizing the object, our observers had to segment it (mark its parts or boundaries) and then retain this information in order to segment a novel object.

While it is clear that observers must accumulate information about object boundaries in order to perform this task, the nature of this information is unclear. To perform this task, observers could have retained information about the entire object, the boundary of the object, or just the boundary between the object and the target. But it is also possible that observers do not retain any information about the objects at all. That is, instead of marking blocks owned by familiar objects and then counting unmarked blocks, observers could have marked the blocks that were not owned by objects and then counted the marked blocks. Of course this latter strategy still requires object recognition to guide object segmentation, since it still requires determining whether a block is owned by a familiar object. [9]

The reader may get a rough idea of the phenomenology of this experiment by trying the following word game. Embedded in this letter string are five, four-letter words and a four-letter nonsense word:

`biketreechipretoitchoven`.

Note first that there are many four letter clusters that are not words but are also not the target because they belong to words (e.g., `ketr` or `eech`). Because there are so many non-target non-words, a strategy of looking for non-words would produce poor performance. A more reliable strategy is to first look for the words (`bike`, `tree`, `chip`, `itch`, `oven`) and then determine which letters are left over (`reto`). While this task obviously differs from ours in many key respects, we suggest that it also requires a kind of recognition-driven segmentation.

This idea of marking an internal scene representation in order to determine spatial relationships was developed in detail by Shimon Ullman [25, 26]. He proposed that after an automatic, bottom-up analysis of the image, a number of task-driven, attentive routines were used to recover spatial relations. We suggest that the top-down segmentation we are exploring in these experiments involves just such a task-driven, attentive process. Attention is required to recognize the objects and it is required to keep track of their boundaries in visual working memory. Our finding that performance drops off sharply when the number of objects exceeds four is consistent with the evidence that four objects is the capacity limit for visual working memory [19, 20, 8, 10].

Our main goal for these experiments was to begin to develop methods for studying what we see as a significant gap between research in mid-level and high-level vision. Researchers interested in mid-level vision have set as their goal, not object segmentation, but surface segmentation. This is because objects are ill-defined as entities in the world. (As Marr wrote, ″Is a nose an object? Is a head one? Is it still one if attached to the body?″ [11].) And similarly objects are ill-defined as entities in the visual image. Surfaces, in contrast, do not suffer from these shortcomings. Surfaces can be defined mathematically and it is thought that they can be recovered directly from an image using only a small set of generic assumptions. Thus the goal of mid-level vision is generally construed as building, from the bottom-up, a surface representation. On the other hand, researchers interested in the high-level problem of object recognition do not start with a surface representation, they start with a segmented object. If mid-level vision ends with segmented surfaces and object recognition begins with segmented objects, how does the visual system span this gap?

One might argue that the gap is only illusory because most objects are delineated by visual cues which makes their recovery straightforward. Certainly, many objects appear to be well-defined by one Gestalt grouping rule or another. But we think this is an illusion similar to the illusion we experience during spoken-language perception. When we listen to someone speaking, we have the impression that there are distinct boundaries that allow us to pick out the words prior to recognition.

[9] While recognizing the limitations of introspection, we think it is worth mentioning that the observers in the second experiment reported using both strategies. When the number of objects was small, these observers thought they identified the boundaries of the objects and then counted the blocks between these boundaries. When the number of objects was large, however, they claimed they could no longer keep track of the objects and so instead tried to systematically scan the scene and count blocks that did not belong to any object.

But in natural speech, word boundaries are not consistently marked by pauses or other stimulus cues. (We realize this when we listen to speakers of a foreign language.) The word segmentation we perceive arises less from cues in the stimulus than from our recognition of the words [3]. Our claim is that there is often an analogous situation in visual perception: the object segmentation we perceive arises not from cues in the stimulus but from our recognition of the objects. Of course, this would suggest that we can only segment familiar objects. The experiments reported here examine one of the ways in which object recognition might assist in the segmentation of an unfamiliar object.

Acknowledgments

We are grateful for support from NSF grant SBR-9729015 (Bravo), and an Alfred P. Sloan Fellowship, NSF CAREER award IIS-99-83806, and NSF grant EIA-98-02068 (Farid).

References

[1] H.B. Barlow. Vision tells you more than 'what is where'. In Andrei Gorea, editor, *Representations of Vision: Trends and Tacit Assumptions in Vision Research*. Cambridge University Press, 1991.

[2] H.G. Barrow and J.M. Tenenbaum. Computational vision. *Proceedings of the IEEE*, 69:572–595, 1981.

[3] M.R. Brent. Speech segmentation & word discovery: a computational perspective. *Trends in Cognitive Science*, 3, 1999.

[4] P. Cavanagh. What's up in top-down processing? In Andrea Gorea, editor, *Representations of Vision*. Cambridge University Press, 1991.

[5] E.C. Freuder. Knowledge-mediated perception. In Eileen Schwab and Howard Nusbaum, editors, *Pattern Recognition by Humans and Machines*. Academic Press, 1986.

[6] K. Gottschaldt. Gestalt factors and repetition. In Willis D. Ellis, editor, *A Source Book of Gestalt Psychology*. Routledge and Kegan Paul Ltd., 1938.

[7] H.L. Kundel and C.F. Nodine. A visual concept shapes image perception. *Radiology*, 146:363–368, 1983.

[8] J. Lachter and M. Hayhoe. Capacity limitations in memory for visual locations. *Perception*, 24:1427–1441, 1995.

[9] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic, 1985.

[10] S.J. Luck and E.K. Vogel. The capacity of working memory for features and conjunctions. *Nature*, 390:279–280, 1997.

[11] D. Marr. *Vision*. W. H. Freeman and Company, 1982.

[12] L.H.M. Mens and E.L.J. Leeuwenberg. Can perceived shape be primed? the autonomy of organization. *Giornale Italiano di Psicologia*, 20:821–836, 1993.

[13] C. Moore and P. Cavanagh. Recovery of 3d volume from 2-tone images of novel objects. *Cognition*, 67:45–71, 1998.

[14] A. Needham and J. Kaufman. Infants' integration of information from different sources in object segregation. *Early Development and Parenting*, 6:137–148, 1997.

[15] M.A. Peterson and B.S. Gibson. The initial identification of figure-ground relationships: Contributions from shape recognition processes. *Bulletin of the Psychonomic Society*, 29:199–202, 1991.

[16] M.A.
Peterson and B.S. Gibson. Shape recognition inputs to figure-ground organization in three-dimensional displays. *Cognitive Psychology*, 25:383–429, 1993.

[17] M.A. Peterson and B.S. Gibson. Must figure-ground organization precede object recognition? an assumption in peril. *Psychological Science*, 5:253–259, 1994.

[18] M.A. Peterson and B.S. Gibson. Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. *Perceptions and Psychophysics*, 56:551–564, 1994.

[19] Z.W. Pylyshyn and R.W. Storm. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, 3:179–197, 1989.

[20] R.A. Rensink. Visual search for change: A probe into the nature of attentional processing. *Visual Cognition*, 7:345–376, 2000.

[21] L.G. Roberts. Machine perception of three-dimensional objects. In Tippet et al., editor, *Optical and Electro-optical Information Processing*. MIT Press, 1966.

[22] A. Selinger and R.C. Nelson. A perceptual grouping heirarchy for appearance-based 3d object recognition. *Computer Vision and Image Understanding*, 76:83–92, 1999.

[23] E.S. Spelke. Principles of object perception. *Cognitive Science*, 14:29–56, 1990.

[24] J. Tanaka, D. Weiskopf, and P. Williams. The role of color in high-level vision. *TRENDS in Cognitive Sciences*, 5:211–215, 2001.

[25] S. Ullman. Visual routines. *Cognition*, 18:97–159, 1984.

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