to appear: **Perception, 2004.**

# Search for a category target in clutter

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

**An airport security worker searching a suitcase for a weapon is engaging in an especially difficult search task: the target is not well-specified, it is not salient and it is not predicted by its context. Under these conditions, search may proceed item-by-item. The experiment reported here tested whether the items for this form of search are whole familiar objects. Our displays were composed of color photographs of ordinary objects that were either uniform in color and texture (simple), or had two or more parts with different colors or textures (compound). The observer's task was to detect the presence of a target belonging to a broad category (food). We found that when the objects were presented in a sparse array, search times to find the target were similar for displays composed of simple and compound objects. But when the same objects were presented as dense clutter, search functions were steeper for displays composed of compound objects. We attribute this difference to the difficulty of segmenting compound objects in clutter: compared with simple objects, bottom-up grouping processes are less likely to organize compound objects into a single item. Our results indicate that while search rates in a sparse display may be determined by the number of objects, search rates in clutter are also affected by the number of object parts.**

Visual search Clutter Object segmentation

[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

The visual search paradigm has been an immensely popular tool for studying the efficiency of visual processing (Neisser, 1967; Atkinson, Holmgren, & Juola, 1969; Treisman & Gelade, 1980; Treisman, 1988; Egeth & Yantis, 1997; Wolfe, 1998). Because the purpose of these experimenters is to isolate and study a particular stimulus feature or a particular psychological mechanism, search stimuli have traditionally been very simple and highly artificial. In addition to serving as a research tool, the visual search paradigm also has practical applications. We engage in visual search numerous times each day (for our car, keys, glasses), and visual search is a central task in some professions (baggage screening, X-ray reading). To understand everyday vision or to improve task performance, it would be useful to extend visual search research to these real-world tasks. Recently, a number of experimenters have begun to study visual search using photographs of real objects and real scenes (Zelinksy, Rao, Hayhoe, & Ballard, 1997; Moores, Laiti, & Chelazzi, 2003).

Most visual search research focuses on the task of detecting a specific target object. In these experiments, the same target is used on every trial, or, alternatively, an image of the search target is shown to the observer prior to the search stimulus. In both cases, there is no uncertainty, not even lighting or viewpoint uncertainty, as to the target's appearance. These experiments have led to the hypothesis that observers maintain an image of the search target in working memory and use this image as a ”search template” (Duncan & Humphreys, 1989; Rao, Zelinsky, Hayhoe, & Ballard, 2002). This template is matched, in a parallel process, against the search stimulus. Neurophysiological experiments in which monkeys perform a comparable task have been interpreted in a similar way. These experiments indicate that top-down inputs from working memory enhance the activity of extrastriate neurons selective for the target stimulus (Chelazzi, Duncan, Miller, & Desimone, 1998). These top-down effects can be very selective because the representation of the target in working memory is very specific.

While a search template implemented through selective top-down enhancement might explain search for a specified target, this kind of parallel process seems less suited to search for a category target. This is the task faced, for example, by an airport security worker searching baggage for potential weapons. In this case, the target could have a variety of colors, shapes, and sizes, and it could even be an object that the observer has never seen before. It is difficult to imagine how a search template could incorporate this high degree of variance and still be selective for the target. Instead of simple template or feature matching, search for category targets would seem to require that at least some regions of the display be processed to a deeper level.

The variability of the target's appearance is not the only difficulty faced by the airport security worker. The high degree of clutter and the lack of predictive context also pose a problem. Under some conditions, category targets can be detected very rapidly, but this has only been demonstrated when the background clutter is minimal or the context is predictive (VanRullen & Thorpe, 2001; Li & R. Van Rullen, 2002), but also see (Johnson & Olshausen, 2003). These experiments showing rapid categorization have used professional photographs in which the target is the subject. Thus, the target is generally centered in the frame, high contrast, unoccluded, and in a typical setting. This last characteristic might be especially important because there is some evidence that rapid scene categorization may be crucial for rapid target categorization (Torralba & Oliva, 2003; Torralba, 2003). We would argue then that the reports of rapid target categorization do not necessarily indicate that it is possible to simultaneously recognize (or categorize) all of the objects in a cluttered scene. In fact, we assume that the opposite is true: that the recognition and categorization

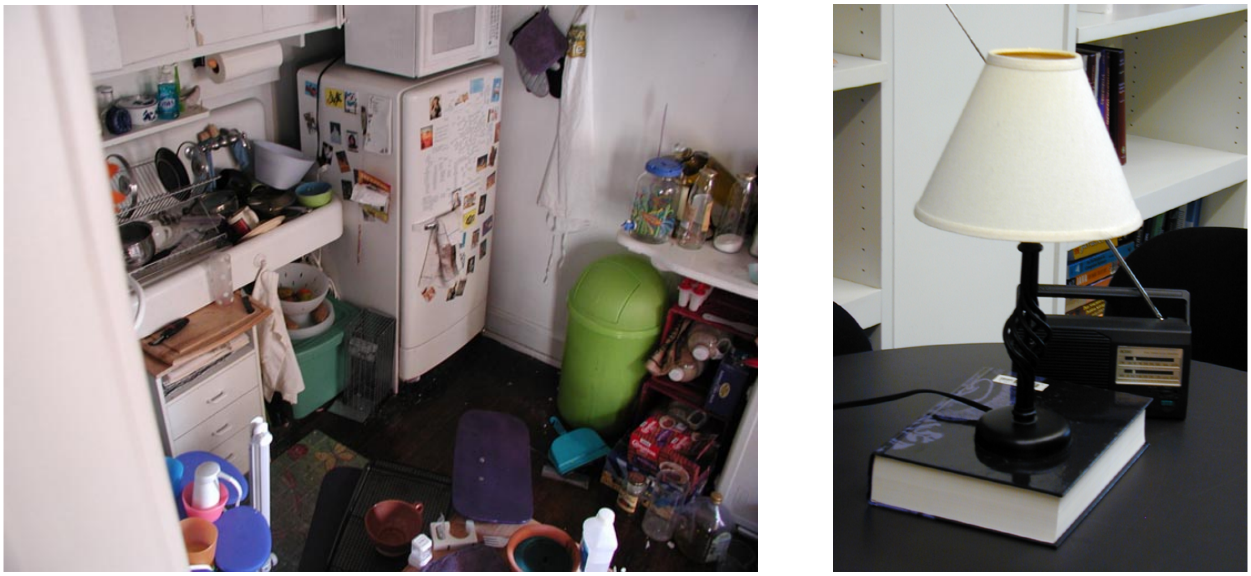

**Figure 1:** Cluttered natural images. It is unlikely that preattentive grouping processes could organize the image on the right into a lamp, a book and a radio.

of multiple objects involves a serial process. Our assumption is based in part on the well-known theoretical argument that the highest levels of visual processing are likely to involve a distributed feature representation. With such a representation, an ambiguity arises when numerous objects are represented simultaneously (von der Malsburg, 1999). Thus, one role of selective attention may be to limit the number of objects that are recognized at any given moment (Moran & Desimone, 1985; Olshausen, Anderson, & VanEssen, 1993; Treisman, 1999), but also see (Riesenhuber & Poggio, 1999; Ghose & Maunsell, 1999).

In this paper we are interested in examining search for a category target in clutter. The important characteristics of this task are that the target is not known, the target is not salient, and the target's existence and location are not predicted by context. Under these stringent conditions, we assume that observers resort to an item-by-item search. Our specific aim is to characterize the items of this search. There is a prevalent, but often implicit, assumption that these items are familiar objects. The notion is that bottom-up grouping processes organize the scene into objects and that visual attention then selects a small number of objects for recognition (see for example, (Olson, 2001)). While we agree with the idea that visual attention may select items for recognition, we doubt that these items correspond to familiar objects. This is because, in clutter, it may be impossible to segment, bottom-up, whole familiar objects. Consider, for example, an object made from two different materials (e.g., a table lamp with a linen shade and a metal base, Figure 1). When this object is juxtaposed with other objects, the within-object boundaries may be as salient as the between-object boundaries (Spelke, 1990). Similarly, occlusions pose a problem for segmentation because they may cause an object to be visible only in fragments. Observers may still group these fragments using color or texture similarity if the object is made from a single material. [3] If the

[3] Of course, even if the object has uniform color and texture, illumination and projection effects will inevitably cause

object has multiple parts made from different materials, however, it seems less likely that preattentive processes will reliably group the object's fragments. In such cases, it may be necessary to recognize these objects in order to accurately segment them from their background. Thus, when observers search a cluttered scene for a category target, we question whether they can select whole objects.

To test the idea that observers select and reject whole objects, we used two kinds stimulus arrangements. In one, photographs of familiar objects were placed in a sparse array. The objects were clearly separated, even when viewed peripherally. This sparse arrangement is typical of traditional search experiments. In the other arrangement, the familiar objects were positioned randomly on the computer screen. This random arrangement, while clearly not typical of most natural scenes, mimics the dense clutter that one might see in a suitcase, a kitchen drawer or a toy chest. In our experiment, observers searched for a food target. Because these targets were drawn from a diverse category and so varied in color, shape and texture, observers could not use the efficient strategy of searching for a distinctive feature or set of features. Instead, we assumed observers would resort to an item-by-item search. The question, then, is whether search times would increase with the number of objects or with the number of object parts. (We are using the term ”part” in non-standard way: here, parts must differ in their color or texture.) To test this, we generated cluttered displays composed of simple objects (e.g., a wooden stool) or compound objects (e.g., a paint roller). If these familiar objects are the items for search, then we would expect that the time it takes observers to find the target would increase with the number of objects, regardless of the object type. On the other hand, if preattentive segmentation does not always yield whole objects, especially when these objects are composed of two materials, then we would expect search times to increase more rapidly for compound objects than for simple objects.

## 2 Stimuli

Our stimuli were composed of color photographs of familiar objects. We hand-selected from the Hemera photo-object collection (www.hemera.com) 132 distractor objects, Figure 2, and 44 target objects, Figure 3. Half of the distractor and target objects were selected because they were composed of one material. That is, they appeared to have largely uniform reflectance and surface texture (e.g., a cucumber, a fire hydrant). We refer to these as ”simple” objects. The other half of the objects were selected because they had two parts with obviously different colors or textures (e.g., a pineapple, a hand trowel). These ”compound” objects were not simply multi-colored; the different colors had to correspond to different parts, and so a striped sock would not be considered compound object.

Displays were composed of 6, 12, or 24 objects. In half the displays, one of these objects was a food target. The distractor objects were selected randomly and without replacement from either the simple or compound sets. Thus, the distractors in a given display were different examples of a single object type. The target object, when it appeared, was selected from either set. Thus, the same targets were used in both the compound displays and the simple displays.

The objects depicted in the images ranged in size from a garbage can to a car key, but the images themselves were all scaled to have the same area ($16,000$ pixels). Before an image was added to

color and texture variation in the image of the object. This may make grouping such fragments non-trivial, but at least it is plausible.

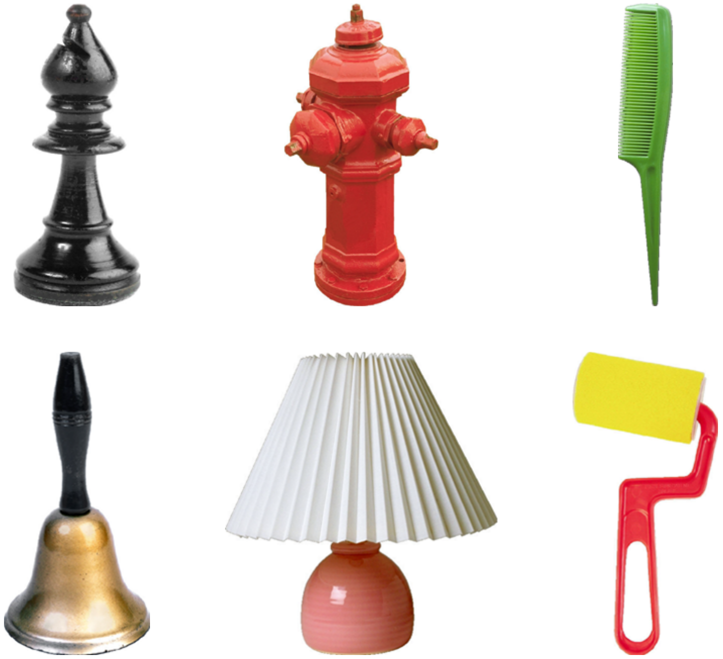

**Figure 2:** Examples of simple (top) and compound (bottom) distractors.

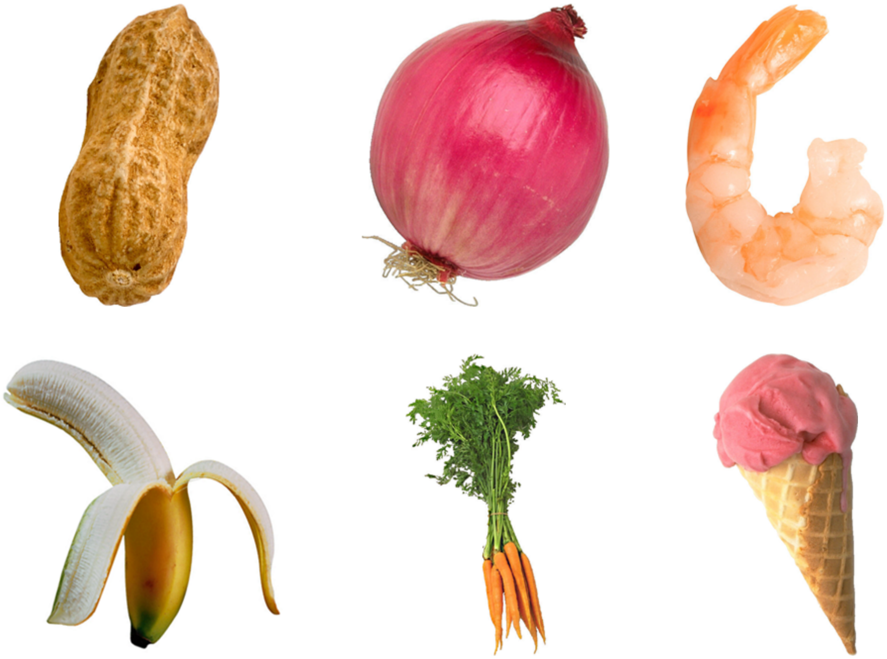

**Figure 3:** Examples of simple (top) and compound (bottom) targets.

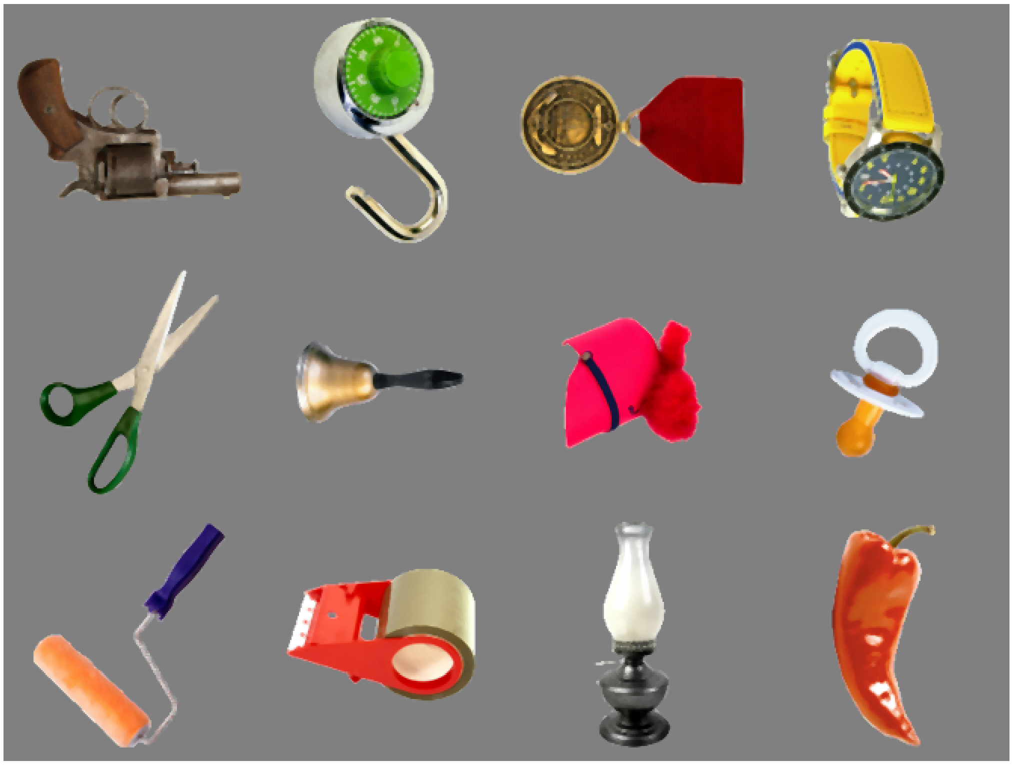

**Figure 4:** A sparse display composed of compound objects and a simple food target.

a display, its area was rescaled by a random factor between $1.0$ and $0.5$ and its orientation was rotated by 0, 90, 180 or 270 degrees.

In addition to using two kinds of objects, we also used two kinds of object arrangements. In the sparse arrangement, the objects were positioned on a uniform array of size $2 \times 3$, $3 \times 4$, or $4 \times 6$, Figure 4. The average width and height of an object was 120 pixels (about $3$ degrees of visual angle), and the average space between objects was 60 pixels. In the clutter arrangement, the objects were positioned randomly as in Figure 5. The area of the display was scaled with the number of objects to keep the average overlap between objects fixed at $20\%$. If more than $40\%$ an object was obscured, the display was recreated. Each target had an average occlusion of $20\%$ for both the composed and simple displays.

The stimulus generation required both manual and automated steps. First, anti-aliased masks were created for each image. This manual step was done in Adobe Photoshop. Then the masks and images were randomly scaled, rotated and positioned in each stimulus. This automated step was done in MatLab. The stimuli were generated off-line, and a new stimulus was generated for each of the observer's 1,056 trials. The stimuli were displayed on an Apple PowerBookG4 using MatLab and PsychToolbox routines (Brainard, 1997; Pelli, 1997).

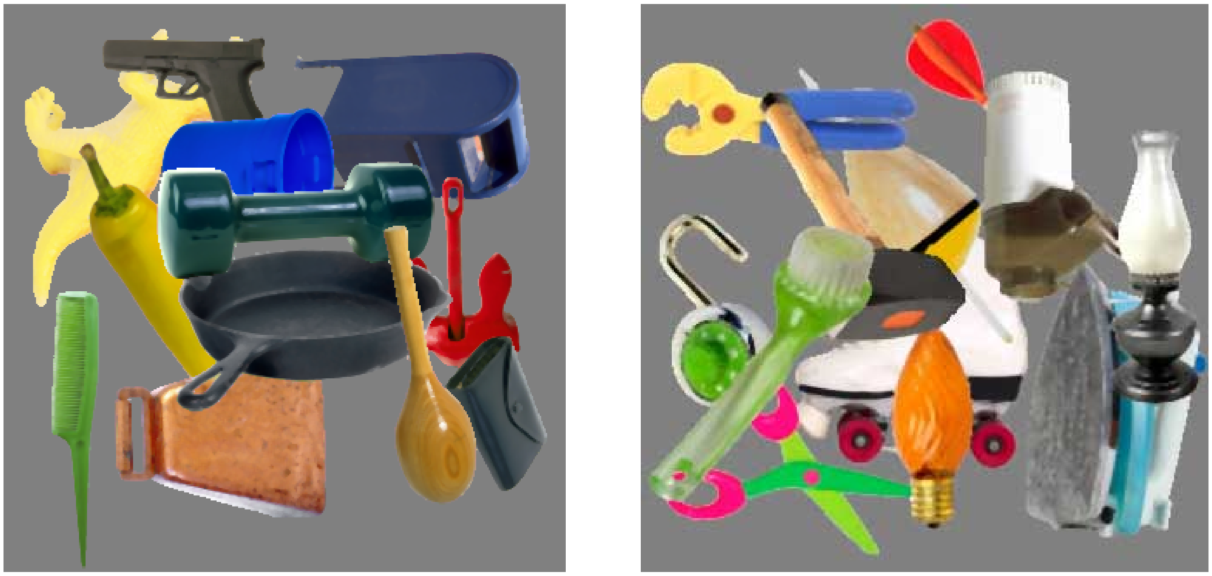

**Figure 5:** Cluttered displays composed of simple (left) and compound (right) objects. A target is present on the left but not on the right.

### 2.1 Procedure

In all, there were 24 conditions: two arrangements (sparse or clutter); two distractors types (simple or compound), three levels of distractor number (6, 12 or 24) and target present or absent. The two arrangement conditions were varied across observers. All other conditions were varied within observers and were completely intermixed within blocks of trials. Each observer participated in two sessions within a one week period. During these sessions, the observers ran a total of 22 blocks of 48 trials each.

The observer initiated the first trial in each block, and the stimulus remained on until the observer responded by pressing either the 8 key (food absent) or the 9 key (food present). Auditory feedback was given after incorrect responses. The next stimulus was presented after a one second delay. The first two trials of each block were discarded as practice.

### 2.2 Participants

The observers were fourteen Rutgers-Camden students who participated to fulfill a course requirement for Introductory Psychology. None of the observers were aware of the purpose of the study.

## 3 Results

We analyzed the data separately for the two target types and found no significant difference. So within each condition, the data from simple target and compound target trials were pooled.

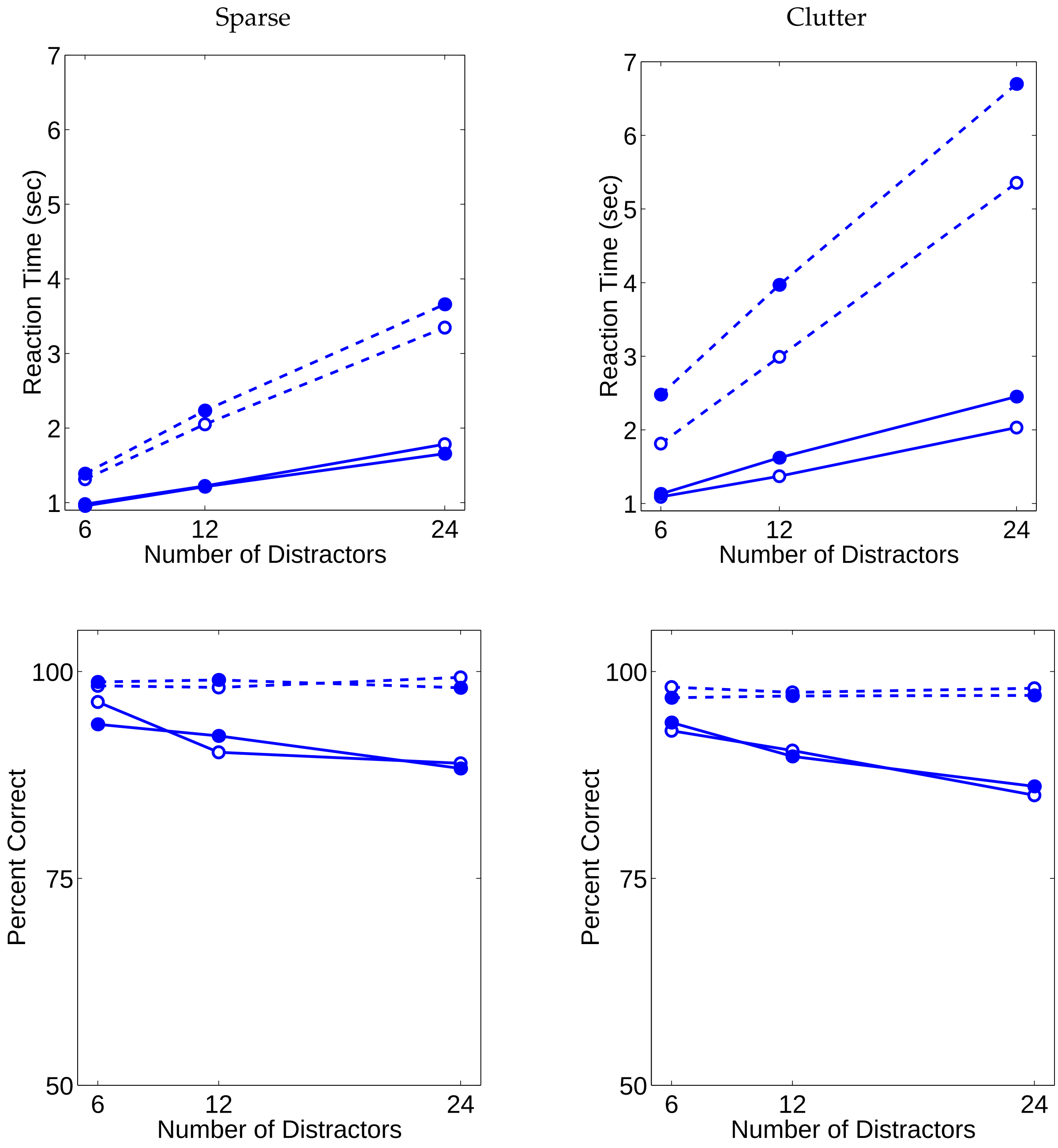

**Figure 6:** Averaged reaction times (top) and percent correct (bottom) as a function of the number of objects in the display. Open circles correspond to simple objects, filled circles to compound objects. Dashed lines correspond to target absent trials, solid lines to target present trials. See Figure 7 for the slope and intercepts.

|  | Sparse |  | Clutter |  |
|---|---|---|---|---|
|  | present | absent | present | absent |
| simple | 45/700 | 112/666 | 53/760 | 197/631 |
| compound | 39/734 | 125/676 | 73/716 | 233/1,114 |
| significance | -/- | */- | */- | */* |

**Figure 7:** Slopes (msec/item)/intercepts (msec) from Figure 6. The asterisks in the bottom row indicate significant differences between the parameters for simple and compound objects (paired t-test, $p < 0.05$).

We also analyzed the data separately for the two experimental sessions run by each observer. Although observers were generally faster during the second session, they produced the same pattern of results on both days. These data were pooled as well.

The average results for the seven observers in the sparse condition are shown on the left side of Figure 6. The open circles correspond to simple object displays, the filled-circles to compound object displays. The solid lines indicate that the target was present, the dashed lines indicate that the target was absent. These reaction time functions were well fit by lines, the slopes and intercepts of which are given in Figure 7. Since we are interested in whether simple and compound objects produce similar results, we used a paired t-test to compare the slope and intercept values for the search functions corresponding to these two object types. When the target was present in the sparse condition, we found no difference in performance between displays composed of simple objects and those composed of compound objects. When the target was absent, the slope for the compound objects was slightly, but significantly, greater than that for the simple objects.

The average results for the seven observers in the clutter condition are shown on the right side of Figure 6. Here we see a more noticeable difference in the reaction times for displays composed of simple objects (open circles) and compound objects (filled circles). Again, the data were well fit by lines and the slope and intercepts of these lines are given in Figure 7. When the target was present, the slope for the compound objects were steeper than that for the simple objects. When the target was absent, the slope and intercept for the compound objects were greater than those for the simple objects.

A MANOVA of the target-present data revealed main effects for set size (F$[2, 24] = 79.6$, p $= 0.00$) and object type (F$[1, 12] = 8.32$, p $= 0.014$). The main effect of display arrangement did not reach significance (F$[1, 12] = 4.64$, p $= 0.052$). All two-way interactions were significant, as was the three-way interaction (F$[2.24] = 5.64$, p $= 0.009$). This last interaction bears directly on our hypothesis. We predicted that the set-size effect would be similar for compound and simple objects when these objects were sparsely arranged, but not when they were were randomly arranged.

# 4 Discussion

To briefly summarize our results, we found that when the compound and simple objects were arranged in a sparse array, they produced similar search times. In particular, we found no difference in performance when the target was present and only a small difference when the target was absent. When the objects were arranged as dense clutter, however, a clear difference between the two types of objects emerged. Observers were slower to determine the target's presence amongst compound objects, and they were extremely slow to determine the target's absence in displays composed of compound objects.

The critical difference between the sparse and clutter arrangements is in the difficulty they pose for object segmentation. In sparse displays, the objects are effectively ”pre-segmented”, and so, we argue, these displays measure the time it takes observers to judge whether a whole object is a food target. Because search times were similar for the simple and compound objects in sparse arrays, these objects appear to be similarly discriminable from food. In clutter displays, however, the observers must segment the objects. Thus, the different pattern of results for these displays can reasonably be attributed to differences in object segmentation. The finding that, in clutter, search times were considerably slower for the compound objects than for the simple objects suggests that observers have greater difficulty segmenting compound objects. In particular, we would argue that when these objects appear in clutter, their parts cannot always be grouped through bottom-up processes.

This failure to group object parts could cause observers to treat each part of the object as a separate item which they must select and reject independently. To illustrate this idea, consider an observer searching for food in a display with a partially occluded table lamp. The observer might select an image chunk corresponding to the lamp base and reject it as not being food. This rejection need not involve recognition, it could be based on the decision that the item is made from some inedible material like wood or metal, or it could be that the item simply doesn't look like any familiar food. Later, the observer might independently select a chunk corresponding to the lamp shade and reject it. Since our compound objects had at least twice as many parts as our simple objects, one might expect search slopes for the compound displays to be at least twice that for simple displays. But the parts within compound objects were not randomly arranged, and so grouping cues like colinearity sometimes allowed the preattentive segmentation of whole objects. In addition, because the objects themselves were randomly positioned, some objects were not obscured by clutter. [4]

An alternative explanation for our result is that observers select and reject compound objects in their entirety, but they do so only after recognition-driven segmentation. Consider again an observer searching for food in a display with a table lamp. By this second account, the observer might select the base and recognize it as the bottom of a lamp. This would cue the observer to select the shade and then reject the entire lamp. In this case, there are two selection steps, one involving the results of image-driven grouping processes, the other involving the results of recognition-driven grouping processes. Thus, the items of the item-by-item search would be whole objects, but the amount of processing required to select and reject whole objects would depend on the complexity of the object. We have some preliminary data suggesting that under the conditions of our experiment, observers use the former strategy. But regardless of which account best describes this process, our results show that clutter can have a significant effect on the processes underlying visual search.

[4] The idea that search rates depend upon the number of object parts might also suggest that search for compound targets will be faster than search for simple targets. (Compound objects, having twice as many parts as simple objects, would seem to provide twice as many targets.) But this reasoning assumes that the observer has two separate chances of landing on the target. If observers use an orderly scan path, then given the close spatial proximity of the target parts, the chances of landing on each of the target parts would be highly correlated.

We wish to emphasize again that our task of searching for a category target in clutter differs in a fundamental way from the search for a known target. When the observer knows the size, color, shape or likely location of the target, the observer can use a search template or specific target feature to guide search. Only those stimuli that share this feature would be given further processing (Cave & Wolfe, 1990; Egeth, Virzi, & Garbart, 1984; Nakayama & Silverman, 1986; Treisman & Sato, 1990; Folk, Remington, & Johnston, 1992; Rossi & Paradiso, 1995). It is important to keep this distinction in mind when contrasting our experiment with another experiment that examined the effect of clutter on search (Wolfe, Oliva, Horowitz, Butcher, & Bompas, 2002). In this earlier experiment, the observer searched for a known target (a yellow "T" ) amongst distractors that closely resembled the target (yellow "L"s). These letter stimuli were superimposed on either a blank background or a cluttered scene. The clutter caused an increase in the intercept but not the slope of the search function. The authors interpreted this as evidence that observers could "segment out" the clutter in a single step. Presumably, observers eliminated the clutter by selecting only those items that possessed the distinctive target feature (e.g., yellow). In our experiment, the targets had no common distinctive feature and so observers could not use the absence this feature to eliminate the clutter.

We undertook this experiment because we questioned the view that visual attention selects whole objects prior to recognition. We were skeptical of this view because we believe that bottom-up grouping processes cannot always extract whole objects from cluttered scenes. In particular, when an object's parts are made from different materials, it may be impossible to group these parts through a purely bottom-up process. The results of this experiment support this idea: observers were slower to find targets in scenes composed of compound objects than scenes composed of simple objects. We have offered two explanations for this result: observer may treat compound objects as multiple items, or they may segment these objects with a recognition-driven process. Either way, these results suggests the need to modify simple accounts of segmentation, attention and recognition.

## Acknowledgments

We are grateful for the insightful comments of two anonymous reviewers. 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).

## 5 References

Atkinson, R., Holmgren, R., & Juola, J. (1969). Processing time as influenced by the number of elements in a visual display. *Perception and Psychophysics*, *6*, 321–326.

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

Cave, K., & Wolfe, J. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology*, *22*, 225–271.

Chelazzi, L., Duncan, J., Miller, E., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory guided visual search. *Journal of Neurophysiology*, *80*, 2918–2940.

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*, 433–548.

Egeth, H., Virzi, R., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 32–39.

Egeth, H., & Yantis, S. (1997). Visual attention: control, representation and time course. *Annual Review of Psychology*, *48*, 269–297.

Folk, C., Remington, R., & Johnston, J. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 1030–1044.

Ghose, G., & Maunsell, J. (1999). Specialized representations in visual cortex: a role for binding? *Neuron*, *24*, 79–85.

Johnson, J., & Olshausen, B. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision*, *3*, 499–512.

Li, F., & R. Van Rullen, C. Koch, P. P. (2002). Rapid natrual scene categorization in the near absence of attention. *Proceedings of the National Academy of Science*, *99*, 8378–8383.

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience*, *6*, 182–189.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *4715*, 782–784.

Nakayama, K., & Silverman, G. (1986). Serial and parallel processing of visual feature conjunctions. *Nature*, *320*, 264–265.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton, Century, Crofts.

Olshausen, B., Anderson, C., & VanEssen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*, 4700–4719.

Olson, C. (2001). Object-based vision and attention in primates. *Current Opinion in Neuroscience*, *11*, 171–179.

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

Rao, R., Zelinsky, G., Hayhoe, M., & Ballard, D. (2002). Eye movements in iconic visual search. *Vision Research*, *42*, 1447–1463.

Riesenhuber, M., & Poggio, T. (1999). Are cortical models really bound by the binding problem? *Neuron*, *24*, 87–93.

Rossi, A., & Paradiso, M. (1995). Feature-specific effects of selective visual attention. *Vision Research*, *35*, 621–634.

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

Torralba, A. (2003). Contextual priming for object recognition. *International Journal of Computer Vision*, *53*, 169–191.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.

Treisman, A. (1988). Features and objects. *The Quarterly Journal of Experimental Psychology*, *40A*, 201–237.

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron*, *24*, 105–110.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 459–478.

VanRullen, R., & Thorpe, S. (2001). Is it a bird? is it a plane? ultra-rapid visual categorization of natural and artificial objects. *Perception*, *30*, 655–668.

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, *24*, 94–104.

Wolfe, J. (1998). *Attention*, Chap. Visual search. Psychology Press.

Wolfe, J., Oliva, A., Horowitz, T., Butcher, S., & Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research*, *42*, 2985–3004.

Zelinksy, G., Rao, R., Hayhoe, M., & Ballard, D. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*, *8*, 448–453.