Hybrid visual and memory search for scenes and objects with variable viewpoints

Bochao Zou School of Computer and Communication Engineering, University of Science and Technology Beijing, China

Zhe Huang Zooplus SE, Munich, Germany

Abla Alaoui-Soce Department of Psychology, Princeton University, Princeton, NJ, USA

Jeremy M. Wolfe Visual Attention Lab, Harvard Medical School and Brigham & Women's Hospital, Boston, MA, USA

In hybrid search, observers search visual arrays for any of several target types held in memory. The key finding in hybrid search is that response times (RTs) increase as a linear function of the number of items in a display (visual set size), but RTs increase linearly with the *log* of the memory set size. Previous experiments have shown this result for specific targets (find exactly this picture of a boot on a blank background) and for broad categorical targets (find any animal). Arguably, these are rather unnatural situations. In the real world, objects are parts of scenes and are seen from multiple viewpoints. The present experiments generalize the hybrid search findings to scenes (Experiment 1) and multiple viewpoints (Experiment 2). The results replicated the basic pattern of hybrid search results: RTs increased logarithmically with the number of scene photos/categories held in memory. Experiment 3 controls the experiment for which viewpoints were seen in an initial learning phase. The results replicate the findings of Experiment 2. Experiment 4 compares hybrid search for specific viewpoints, variable viewpoints, and categorical targets. Search difficulty increases from specific viewpoints to variable viewpoints and then to categorical targets. The results of the four experiments show the generality of logarithmic search through memory in hybrid search.

Introduction

In standard visual search, observers search for a target in visual displays containing distractor items (Wolfe, 2020). In "hybrid search," observers search the visual display (e.g., the shelves in the supermarket) for a set of possible targets held in memory (e.g., your shopping list). This is known as "hybrid search" because it combines visual and memory search (Schneider & Shiffrin, 1977; Wolfe, 2012). It is a typical type of search task encountered in daily life. To investigate hybrid visual and memory search behavior in the laboratory, Wolfe (2012) asked human observers to memorize 1, 2, 4, 8, 16, or even 100 objects (memory set) prior to search. To confirm that these objects had been firmly stored in their memory, observers completed a simple "old" or "new" memory recognition test. Next, the observers performed repeated trials of visual search through displays consisting of either 1, 2, 4, 8, or 16 photographs of objects (visual set). The observers' task was to identify if one of the objects in the memory set was present in the search display. The results showed that the search response times (RTs) were a linear function of the number of objects displayed in visual search and a logarithmic function of the number of objects held in memory. Cunningham and Wolfe (2014) offer a model that interprets the basic mechanism of the interaction between visual and memory search: An object in the visual display is selected. In the Wolfe (2012) experiment, with a diverse set of target objects, this visual selection will be essentially random (Wolfe, 2021). That selected item will be compared against the set of target objects held in memory ("memory search"). If it does not match any of them, a new item will be selected. This process will repeat until the selected object matches one of the targets or the search is terminated with a "target absent" response. The time required for each memory search will be a log function of the number of items held in memory. Why is the function logarithmic? One appealing thought is that search through memory is like the child's game of guessing a number between 1 and N. A young

Citation: Zou, B., Huang, Z., Alaoui-Soce, A., & Wolfe, J. M. (2024). Hybrid visual and memory search for scenes and objects with variable viewpoints. *Journal of Vision*, *24*(1):5, 1–16, https://doi.org/10.1167/jov.24.1.5.

https://doi.org/10.1167/jov.24.1.5 Received May 5, 2023; published January 10, 2024 ISSN 1534-7362

child will ask, "Is it 1? Is it 2?" and so on; reaching the correct answer in an average of $(N + 1)/2$ steps. A wiser child will learn a set partitioning strategy: "Is it bigger than $N/2$? If no, is it bigger than $N/4$?" and so on, a strategy that requires $\log2(N)$ steps on average. It is difficult to see how that would be implemented in a memory search. A more plausible hypothesis sees the logarithmic function as a by-product of the mechanics of Ratcliff's diffusion model of recognition (Ratcliff, 1978) or related models. In these models, information about an item accumulates at some rate toward an identification threshold. That threshold should be set to a level that allows for recognition as quickly as possible but not so quickly that a noisy accumulation process will produce a false-positive (false alarm) response. If the target can be any of N items in a memory set, one can imagine N diffusion processes accumulating information. The chance of a false positive will go up because each diffuser has some chance of producing a false positive. Accordingly, the recognition threshold should be moved to a higher level if one wishes to avoid an increase in false positives. At a constant average rate of information accumulation, it takes longer to reach a higher threshold. Leite and Ratcliff (2010) have shown that, if the false-positive rate is held constant, response times will increase logarithmically with the number of diffusers. This seems like a plausible account for the basic hybrid search results.

The basic RT pattern of hybrid search has been replicated in a variety of follow-up experiments. Boettcher and Wolfe (2015) used words instead of objects to comprise the visual and memory set; the results showed that the RTs still increased linearly with the number of the words presented in the visual search array and logarithmically (albeit noisily) with the length of the word list memorized by the observers prior to search. Interestingly, even though the memory set was a grammatically and syntactically well-formed phrase with a clear word order (e.g., "London Bridge is falling down"), search RTs were still logarithmically related to the size of the memory set. No reliable serial-position memory effects of words order were apparent. By manipulating the frequency and recency of object items' appearance in the task, Wolfe et al. (2015) demonstrated that the relative familiarity of the targets and distractors did not influence observers' performance on hybrid search. Cunningham and Wolfe (2014) used object categories (e.g., plants, cars, animals, clothing), and observers searched photographic representations of specific objects looking for members of any of the object categories. This categorical search was more difficult than a search for a set of specific objects, so the RTs were overall longer than the equivalent conditions using specific items in Wolfe (2012). Even so, memory search remained logarithmic, and visual search remained linear. This finding indicates that the hybrid search rule is not dependent on the use of specific items.

Other research has investigated the role of memory in hybrid search. Drew, Boettcher, and Wolfe (2016) conducted a series of experiments to test the hypothesis that visual working memory hosted the search templates of the memory set in hybrid search. They asked participants to perform hybrid search and visual working memory tasks at the same time. For example, trials of an ongoing hybrid search were interleaved with trials of a working memory change detection task. Drew et al. (2016) found the additional visual working memory tasks produced little or no interference on the hybrid search. They concluded that the search templates do not reside in visual working memory. In any case, given that working memory capacity is profoundly limited (7 ± 2 or 4 ± 2; Cowan, 2001), it would be implausible that the large numbers of targets (100 in Wolfe's [2012] second experiment) could reside in working memory. Drew et al. (2016) proposed that the memory set of hybrid search is held in an activated long-term memory (ALTM). ALTM is defined as a portion of long-term memory that can be relevant to the current task (Cowan, 1988).

In the real world, stimuli are not generally seen in just one pose or isolated on a uniform ground. We typically encounter stimuli in continuous scenes and from multiple viewpoints. If, as discussed before, the logarithmic RT functions of hybrid search arise from the mechanics of a diffusion model, one could imagine that less constrained stimuli might not reproduce the standard hybrid results. For instance, consider objects that can vary in their viewpoint. Is a left-facing cow identified by the same diffusion process as the same cow, facing down and to the right? Do different items demand dramatically different numbers of diffusers in order to tolerate variations in viewpoint or is the same fixed cost to allow for viewpoint variation? One could imagine responses to viewpoint variation that would disrupt the standard logarithmic funding. Similarly, identification of scenes might rely on multiple diffusers. Perhaps multiple objects need to be identified in each scene, making a search through scenes something of a hybrid, hybrid search. Again, this could alter the log function.

Here we present the results of hybrid search for scenes (Experiment 1) and viewpoint-varying objects (Experiments 2–4) that indicate the basic hybrid result is robust in the face of these concerns. However, it is that we recognize scenes or deal with viewpoint changes, and those processes continue to generate logarithmic RT $\times$ memory set size functions. Different types of stimuli vary in their mean RT with scenes and viewpoint variable stimuli being slower than simple single objects and faster than broad categories, but the qualitative pattern of results remains similar across stimulus sets. Experiment 1 makes this point for scenes. In Experiment 2, we asked if hybrid search changes if targets can be presented in variable viewpoints

compared to search for targets presented from a single viewpoint. While we found that single-viewpoint search is more efficient than variable-viewpoint search, we continued to find the logarithmic function with a change in memory set size. In Experiment 2, observers were initially exposed to objects in a single pose. In Experiment 3, targets were rotated in the learning phase so that participants were exposed to all viewpoints before starting the search trials. The findings from Experiment 2 were replicated. To further compare object, viewpoint, and category search in a fair manner, Experiment 4 ran an "all-in-one" experiment with a single set of stimuli and found that the search difficulty increased from specific viewpoints to variable viewpoints to categorical targets. Again, while the results of these experiments show that search may be easier or harder with different types of stimuli, the basic pattern of a linear search through the visual display and a logarithmic search through the memory set appear to apply generally.

Experiment 1—hybrid search for scenes

Experiment 1 asks whether hybrid search for scenes obeys the same rules as hybrid search for objects. If hybrid search RTs are based on the number of "objects" in memory, scenes could perform differently, given that there is no good way to count the objects in a scene (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011). Adapting the basic hybrid search paradigm, our experiment had two conditions: scene-specific and scene-categorical. In the scene-specific condition, several specific scene photographs constituted the memory set, serving as the target candidates. In the scene-categorical condition, the memory set was composed of several names of scene categories. The visual display in both conditions consisted of a set of nontarget scenes (distractors) with one target scene always present. The visual set sizes varied from trial to trial. In both conditions, the participants' task was to localize the target scene photo in the display. Moreover, we wanted to compare the search efficiency between these two conditions. Cunningham and Wolfe (2014) compared their object-categorical search data (Experiment 2) with Wolfe's (2012) object-specific data (Experiment 3). They found that the search for one member of a multiple-category memory set was much harder than the search for one member of a set of specific targets. We ask if the situation would be the same for scene stimuli. Is scene-categorical search much harder than scene-specific search? We also compare our scene-specific data to Wolfe's (2012) object-specific data and our scene-categorical data to Cunningham and Wolfe's (2014) object-categorical data. Though there are limitations to comparing across experiments and across observers, these comparisons could tell us whether scene search or object search are qualitatively different.

Methods

Participants

Twelve observers participated in this experiment. They were fluent speakers and readers of English. All of them had either normal or corrected-to-normal vision. Before the experiment, all of them gave informed consent and had passed the Ishihara Color Test. However, due to technical problems with the testing computer, we had to exclude one participant (female, 29) from the data analyses because not all candidate categories were presented during search and certain categories had come up more than once as targets in her scene-categorical search condition. Thus, there were 11 valid data sets for analysis. These data sets were from two male participants and nine female participants. The age range was from 18 to 51 years, with an average of 29.3 ($SD = 8.96$). Sample size was based on Wolfe (2012), where 10 Os were tested. The main hybrid search results are very robust and can be established with relatively modest sample sizes.

Materials and apparatus

For the scene-categorical condition, 15 categories were chosen from the SUN Database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). These categories were amusement park, bathroom, beach, bedroom, bridge, cemetery, city street, kitchen, mountain, restaurant, stadium, staircase, swimming pool, theater, and waterfall. Each category contained a minimum of 150 scene photographs and an average of 175 photographs. In the SUN Database, all scene categories are hierarchically arranged into three levels. Most categories in our experiment were selected from the basic level (the third level). Every category in this level has its clear and unique definition based on WordNet (Fellbaum, 2010; Miller, 1995). For example, "restaurant" is defined as "a room where waiters serve meals to customers." Therefore, every category is distinguishable from one another. Even so, some of the SUN categories are easily confused with each other, for instance, "waterfall block," "waterfall cascade," and "waterfall cataract." In these cases, we created a broader "superordinate" category for them: in this case, "Waterfall." For each of our 15 categories, we screened all photos in each category individually, in order to make certain that no image could be potentially classified into an incorrect category. In the end, we had a total of 2,618 scene photographs as stimuli.

Figure 1. Visual search display. An example for the three visual set sizes.

The smallest category ("amusement park") had 153 exemplars, and the largest ("staircase") had 193.

Each category could be used once and only once as a member of the target memory set in the scene-categorical condition. Thus, no category was used as a target in more than one search block. After the target categories were chosen for the search block, scene photographs from the remaining categories were pooled together and the distractors were randomly picked from that superset. For the scene-specific condition, the same 2,618 scene photographs were used as the stimulus pool, except that they were not treated as members of any categories. Fifteen unique scene photographs were randomly chosen to make up memory target sets for each observer, and distractors were once again randomly picked among the rest. The experimental sessions were coded in MATLAB software using the Psychophysics Toolbox (Brainard, 1997) and carried out on a Macintosh G4 computer. Stimuli were presented on a 20-in. CRT monitor (Mitsubishi Diamond Pro 91TXM). The resolution of the display was set to 1,280 * 960 pixels and an 85 Hz refresh rate. Participants were seated so that their eyes were 57.4 cm from the monitor. At this viewing distance, 1 cm subtended 1° of visual angle.

Both scene-specific and scene-categorical conditions consisted of four search blocks. Each block had one of the four memory set sizes: 1, 2, 4, and 8. Each block consisted of 144 visual search trials. On each trial, two, four, or eight scene photographs appeared as the visual search display. Scene photographs were placed equally distant from each other on an invisible circle with a radius of 10.2 degrees at an approximate viewing distance of 57.4 cm (see Figure 1). Within every block, each visual set size showed up 48 times in random order.

For the scene-specific condition, targets were presented as individual photographs one at a time prior to the start of a block of search trials. For the scene-categorical search condition, target categories were defined by words on the monitor screen, naming each scene category. After the target set was presented, observers needed to pass a memory test before moving on to the visual search. In the memory test, observers judged if a scene image was "old" or "new" by pressing corresponding keys as a response. Scene images from the memory sets were presented one at a time, for 3 s each. Images used as scene exemplars in the category condition and as distractors in either condition during the memory test were not used as stimuli in the subsequent visual search trials. For the scene-specific condition, "old" meant the testing image was one of the targets, and "new" meant it was not, while for the scene-categorical condition, "old" meant that the testing image belonged to any of the target categories in the memory set, and "new" meant it did not. The number of memory test trials was twice the memory set size for that block (50% "old," 50% "new"). Participants need to perform at 90% accuracy or better, twice in a row, in order to pass the memory test and to be qualified for the search task. In practice, the average number of blocks required varied between 2 (the minimum possible) and 3.5 for different memory set sizes for specific scenes and from 2.2 to 3.2 for categorical scenes. As a comparison, the memory test for objects in the original hybrid search study required between 2.7 and 3.4 repetitions for memory set sizes between 2 and 16 (Wolfe, 2012).

Participants performed a localization task on the visual search trials. There was one and only one target on each search trial. Using a localization task, with 100% target-present trials, reduces error rates and thus reduces speed–accuracy trade-offs in a study where RTs are the measure of greatest interest. For instance, Experiment 2 in Cunningham and Wolfe (2014) used this method and produced similar RT patterns to a present/absent hybrid search task but with lower error rates. The participants were asked to make a response by clicking the photograph on the screen with a computer mouse. If the wrong photo was clicked, a warning audio feedback would be given, and this prompted the observers to continue searching until they produced the correct answer. The current trial would not be terminated until the correct

photograph was clicked. The number of any clicks before the correct response was recorded as "bad click" counts for that trial, and the trial was noted as an "error trial." All eight search blocks (four scene-specific plus four scene-categorical) were presented in pseudo-random order, counterbalanced across participants.

Results

In each condition of the current experiment (scene-specific and scene-categorical search), participants were tested on 576 trials, and 11 valid participants' data sets were analyzed. There were 12 different combinations of visual set size (VisSS) and memory set size (MemSS) for each condition. Each combination had 48 trials.

Scene-specific search

To "clean" the data, any trials labeled as "bad clicks" were regarded as error trials and excluded from further analyses. The bad click exclusion rate was 2.9% (182 out of 6,336 trials). The bad click exclusion rate was lowest at VisSS 4/MemSS 2 (1.5%) and highest at VisSS 2/MemSS 4 (3.9%). A two-way repeated-measures analysis of variance (ANOVA) on the arcsine-transformed error rates revealed main effects of visual set size, $F(2, 20) = 5.55$, $p < 0.05$, $\eta_p^2 = 0.36$, but no main effect of memory set size, $F(3, 30) = 0.98$, $p = 0.41$, $\eta_p^2 = 0.09$, or the interaction between memory set size and visual set size, $F(6, 60) = 0.62$, $p = 0.71$, $\eta_p^2 = 0.06$. In addition, trials with RTs less than 200 ms or larger than 4,300 ms (3 *SD* from the mean) were excluded as outliers (0.19% of trials).

We performed a two-way repeated-measures ANOVA on the RTs for the remaining trials. Both of the main effects, VisSS and MemSS, were significant: $F(2, 20) = 268.34$, $p < 0.001$, $\eta_p^2 = 0.96$ for VisSS and $F(3, 30) = 41.48$, $p < 0.001$, $\eta_p^2 = 0.81$ for MemSS. We also found a significant VisSS * MemSS interaction effect, $F(6, 60) = 34.96$, $p < 0.001$, $\eta_p^2 = 0.78$. As we can see in Figure 2, plotting the RTs against VisSS, RT increased linearly for each of the memory set sizes. Figure 3a shows that the effect of MemSS on the RTs seemed to be logarithmic. Each VisSS line in Figure 3a is curvilinear. In order to test the hypothesis that the logarithmic model fits better than a linear model, we performed linear regression tests on both RT * MemSS and RT * log2(MemSS) functions. The test of RT * MemSS produced $R^2 = 0.90$, 0.81, and 0.82 for VisSS of 2, 4, and 8, respectively, while the loglinear regression produced higher R^2 values: 0.99, 0.97, and 0.97. The regression tests suggested a more convincing linear relationship between the RTs and the log2(MemSS)

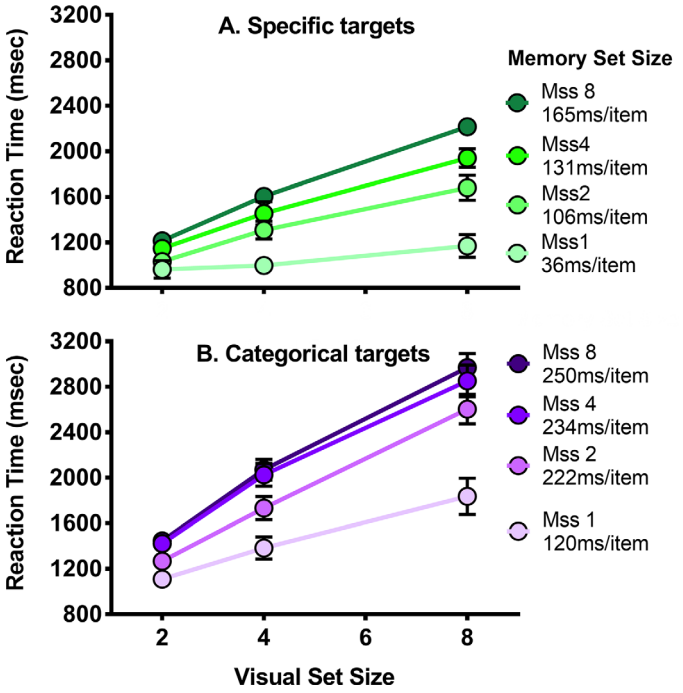


Figure 2. Reaction time as a function of visual set size for the specific (**A**) and categorical scene (**B**) conditions in Experiment 1. Error bars, where visible, are ± 1 *SEM*.

than between RT and MemSS. We also compared the hybrid scene-specific search data with the object-specific data from Wolfe (2012) (Experiment 3, also localization task). The RT × VisSS slopes for scene-specific search were 36 ms/item, 106 ms/item, 129 ms/item, and 165 ms/item for MemSS 1, 2, 4, and 8, respectively. Correspondingly, slopes for specific object search in Wolfe (2012) were 17 ms/item, 49 ms/item, 77 ms/item, and 91 ms/item. It is somewhat perilous to compare across experiments. However, the search for specific objects appears to be about twice as efficient as the search for specific scenes.

Scene-categorical search

As in the specific condition, any response labeled as a "bad click" was regarded as an error trial and excluded from further analyses. This made the error rate 3.3% (212 out of 6,336 trials). A two-way repeated-measures ANOVA on the arcsine-transformed error rates revealed no main effects of visual set size, $F(2, 20) = 2.57$, $p = 0.10$, $\eta_p^2 = 0.20$, or the memory set size, $F(3, 30) = 0.47$, $p = 0.70$, $\eta_p^2 = 0.05$. The error rates in all 12 VisSS/MemSS combinations varied but not in the same manner as in Experiment 2 of Cunningham and Wolfe (2014). In that study, low VisSS/MemSS produced a low error rate while higher VisSS/MemSS produced a higher error rate. Here, the lowest rate of 2.1% was at VisSS 4/MemSS 2 and the highest rate of 4.4% was at

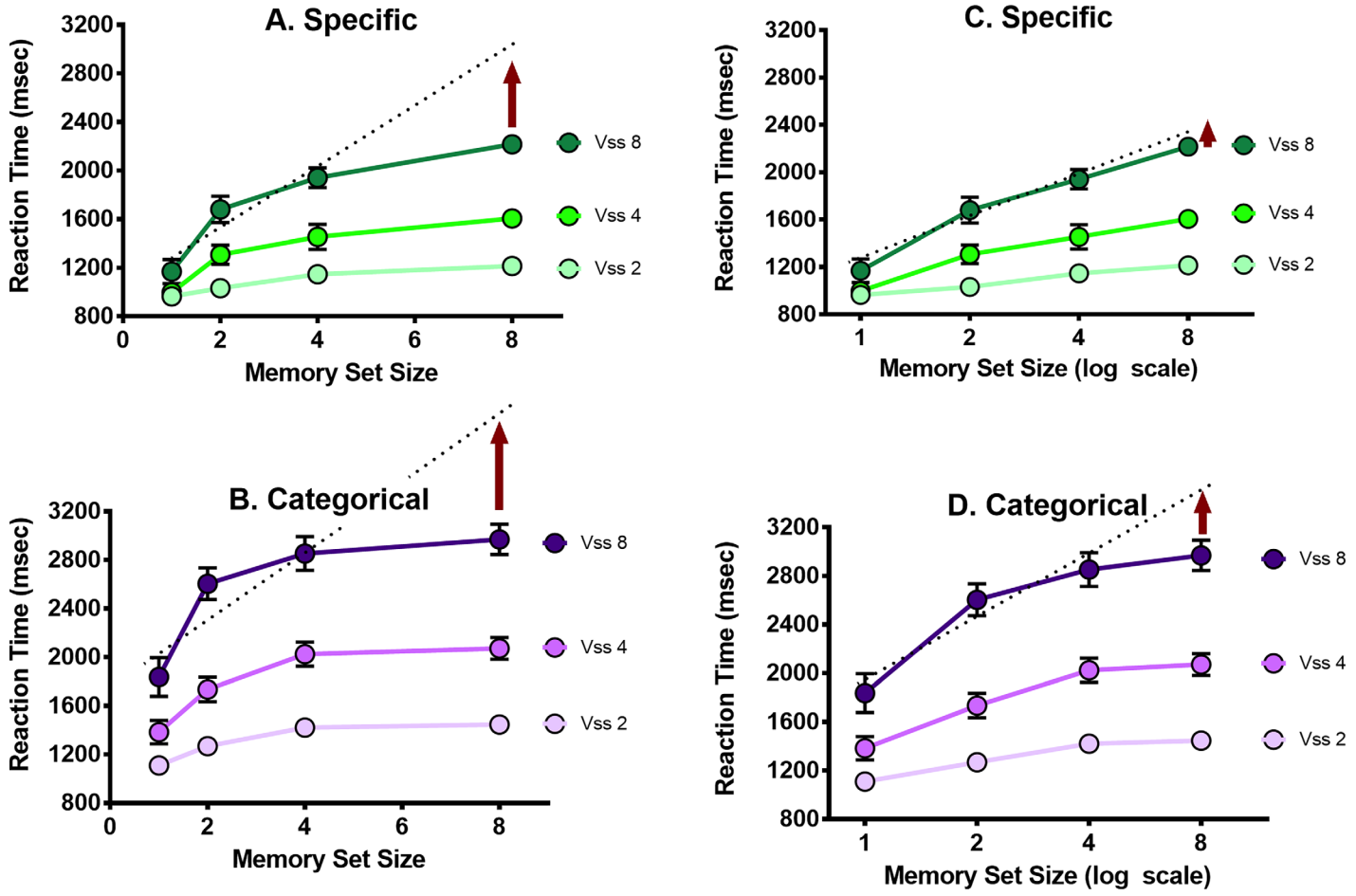


Figure 3. Reaction time as a function of memory set size for the specific (**A**, **C**) and categorical conditions in Experiment 1 (**B**, **D**). A and B show results for Experiment 1 on a linear x-axis. C and D use a logarithmic x-axis. Error bars are ± 1 *SEM*. The dotted line shows a best-fit line through the Vss 8 data for memSS 1, 2, and 4. The red arrow indicates the magnitude of the difference between the extrapolation to MemSS 8 as in Wolfe (2012).

VisSS 8/MemSS 2. The error rate here was much smaller than the object-categorical search of Cunningham and Wolfe (2014). People tended to be faster at finding objects than finding scenes (see below), but they also made more mistakes. In addition, trials with RTs less than 200 ms or larger than 7,300 ms (3 *SD* from the mean) were excluded as outliers (0.68%).

For the remaining trials, a two-way ANOVA on the RTs showed both main effect of VisSS, $F(2, 20) = 236.84$, $p < 0.001$, $\eta_p^2 = 0.96$ (Figure 2B), and main effect of MemSS, $F(3, 30) = 32.05$, $p < 0.001$, $\eta_p^2 = 0.76$ (Figure 3B). There was also a significant interaction between these two, $F(6, 60) = 14.68$, $p < 0.001$, $\eta_p^2 = 0.60$. As we can see by plotting the RTs against VisSS in Figure 2B, each MemSS line increased linearly. The effect of MemSS on the RTs seemed to again produce logarithmic or, at least, curvilinear functions (Figure 3B). In order to compare linear to logarithmic models for MemSS functions, we performed linear regressions on both RT × MemSS and RT × log2(MemSS) functions. The regressions on RT × MemSS produced 0.72, 0.69, and 0.61 R^2 for VisSS 2, 4, and 8, respectively, while for the RT × log2(MemSS), the R^2 reached markedly higher values (0.93, 0.91, and 0.85). As with object categories (Cunningham & Wolfe, 2014), the logarithmical relationship between the MemSS and the RTs is the better fit for the RT × MemSS functions.

We compared the slopes of the VisSS scene category functions with the comparable slopes for object categories in Cunningham and Wolfe (2014), Experiment 2 (note that there was no VisSS 2 in the object-categorical search). The slopes for the Cunningham and Wolfe (2014) object data were 67 ms/item, 90 ms/item, 109 ms/item, and 125 ms/item for MemSS 2 to 8, respectively. The slopes for the scene category data, described here, were 120 ms/item, 222 ms/item, 234 ms/item, and 250 ms/item. As with specific items, categorical object search is about two times more efficient than categorical scene search.

Discussion

The results of Experiment 1 show that hybrid search for specific scene targets produces RTs that are a linear function of the visual set size and an approximately logarithmic function of the memory set size. The same pattern is seen in hybrid search for scene categories. Thus, scenes replicate the pattern of results shown for photographs of specific objects, object categories, alphanumeric characters, and English words or phrases (Boettcher & Wolfe, 2015; Cunningham & Wolfe, 2014; Wolfe, 2012). These results suggest that each scene in the memory set was treated as a single "thing" and not some ill-defined collection of objects for the purposes of identification in the search task. The scene searches were less efficient than the equivalent object searches from previously reported studies. Not too much should be made of this comparison since it could reflect differences in low-level properties of the stimuli (e.g., more self-crowding in the scenes) or high-level factors like some fundamental difference in the recognition of scenes versus objects. It might be hard to disentangle these factors, although some progress might be made by doing an experiment in which Os searched for object targets embedded in scenes. Such a "Where's Waldo" style of hybrid search experiment would seem to require a carefully constructed stimulus set that does not currently exist.

Experiment 2—hybrid search for variable viewpoint

In prior work on hybrid search with objects, observers have searched for either an exact copy of a target object (find this picture of this exact chair) or an object in a target category (find any chair). These conditions could be seen to miss the most common class of real-world object search. In the real world, one would most typically search for a specific object, but one that could be encountered from multiple positions/angles. If you need to find your cellphone, your memory set needs to account for the fact that it can be in an arbitrary orientation in three-dimensional space when you encounter it. These searches for targets that can appear from different viewpoints lie between search for just one view and search for a category. Is variable-viewpoint hybrid search similar to single-viewpoint search? This might be the case if suggesting that search templates are independent of viewpoint (see Biederman & Gerhardstein, 1995; Tarr & Bulthoff, 1995). Alternatively, is variable-viewpoint search like a category search where multiple views behave like multiple instances of a category? To test this, we conducted a hybrid search experiment using specific target objects that could be rendered in multiple viewpoints. We compare this variable-viewpoint condition to a single-viewpoint condition, in which each target appeared in only a single pose.

Methods

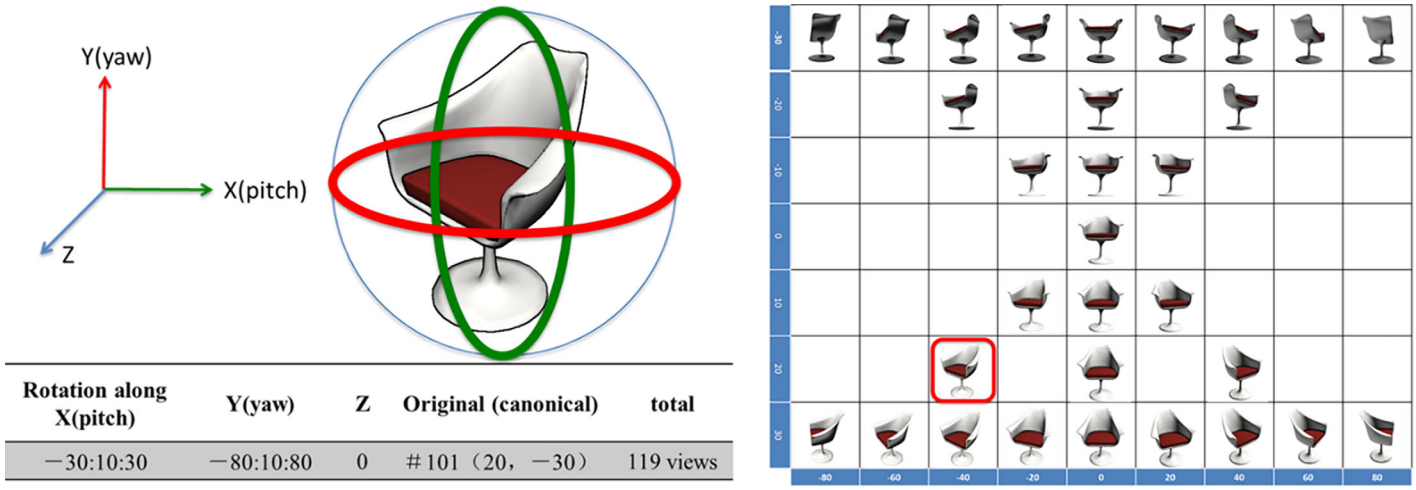
Participants

Fourteen observers (mean age = 26.2 years, SD = 6.7, 11 females) participated in this experiment. All of them had either normal or corrected-to-normal vision. All observers gave informed consent and were paid $10 per hour for their time. Again, sample size was taken from Wolfe (2012) with a few extra observers in case an observer needed to be removed from analysis.

Materials and apparatus

For the viewpoint hybrid search condition, objects with variable viewpoints are needed. We used ShapeNet (Chang et al., 2015), which is a large-scale repository of three-dimensional CAD models of objects, to generate multiple viewpoints for each three-dimensional object model. We rotated each three-dimensional model along the x (pitch angle −30:10:30) and y (−80:10:80) axes, as illustrated in Figure 4. This produced 119 viewpoints for each object. The canonical viewpoint was always designated to be the three-dimensional object model rendered with pitch = 20 degrees and yaw = −30 degrees, relative to a frontal view. ShapeNet Viewer (Version 0.1.0) was used to render different viewpoints of 622 three-dimensional object models. These were selected from 25 categories in ShapeNetCore.v1. The 25 categories were airplane, animal, bag, bed, birdhouse, camera, cap, car, chair, clock, dresser, earphone, guitar, gun, helmet, mailbox, motorcycle, piano, pillow, printer, ship, shoes, sofa, statue, and table. These categories were selected to avoid semantic overlap between categories (e.g., many "toys" can be easily confused with "animals," so that category was excluded). The number of three-dimensional models for each category ranges from 10 to 43. To ensure distinguishability within a category (e.g., a white pillow without any pattern would be a "bad" target/distractor), some three-dimensional models were discarded if they had an excessively plain texture by visual inspection. We have shared the stimuli on Google Drive for reproducibility (https://drive.google.com/file/d/1sxaFCObkkqpQ_8MJ2GRLBqHlKFhDLULM/view?usp=share_link). This process yielded 74,018 images in total. An example is shown in Figure 4.

Experiments were written in MATLAB using the Psychophysics Toolbox (Version 3; Brainard, 1997). Stimuli were presented on a 20-in. monitor with resolution set to 1,280 * 960 pixels and refresh rate set



Rotation along X(pitch)	Y(yaw)	Z	Original (canonical)	total
−30:10:30	−80:10:80	0	#101（20，−30）	119 views

Figure 4. An example of a chair with a variable set of viewpoints generated.

Figure 5. Example stimuli with visual set size 16. The target is a specific chair in any viewpoint; the green rectangle was only shown during feedback.

to 85 Hz. Observers searched visual displays of 4, 8, or 16 photographs of objects for any of 2, 4, 8, or 16 items held in memory. Example stimuli are shown in Figure 5. Every observer was tested on all four memory set sizes over four blocks of trials. At the start of each memory block, the objects in the memory set were presented to the observer. Observers then took a memory test in which they identified pictures of objects as being in or out of the memory set. Observers passed the memory test by scoring over 90% correct. Objects used in the memory test were not used in the subsequent visual search trials. Having passed the memory test, observers searched visual displays where one, and only one, of the items in the display was a target, drawn at random from the different viewpoints created for specific target objects. Distractors were drawn from all of the remaining nontarget sets. Different views of the targets could not appear as distractors. There were 20 practice trials, followed by 300 experimental trials, evenly divided between target-present and target-absent trials and between three visual set sizes: 4, 8, and 16 items. Observers were instructed to indicate whether the target was present or absent with key press as quickly and accurately as possible. The response keys are defined by the Os' choice. At the start of a session, the experiment program asks the participant to press one key for "PRESENT" responses and another key for "ABSENT" responses. The same process was repeated for each of the four memory set sizes. As a replication of Wolfe (2012), we also performed each of the blocks in the experiment using just a single viewpoint for each of the targets in the memory set. We used the present/absent method, rather than the localization method of Experiment 1, because we were replicating

the main experiment of Wolfe (2012). The experiment order was pseudorandomized between four memory set sizes and two conditions (variable vs. single viewpoint) among participants.

Results

The data of participants whose average error rate was more than 20% were excluded from further analysis, leaving a total of 12 participants. The error rates of the two excluded observers were 32.8% and 40.2% (SD = 9.8%), respectively. The average error rate for these 12 participants was 6.8% (range from 5.5%–18.7% with an SD of 3.7%). All error trials were removed from the subsequent RT analysis. In addition, trials with RTs less than 200 ms or larger than 15,000 ms were excluded as outliers (0.12%). A paired sample t test on arcsine-transformed error rates revealed a higher error rate of the viewpoint condition (mean error rate = 8.6%) (miss rate = 10.7%, false alarm = 5.17%) than the single-view condition (mean error rate = 5.9%), $t(11) = -5.4$, $p < 0.001$.

First, we assess the effect of viewpoint on miss error rates and RTs. A paired sample t test on arcsine-transformed error rates shows that the average miss rate for negative pitch angle (−30, −20, −10, $M = 12.0\%$) is higher than that for positive pitch (10, 20, 30, $M = 9.5\%$), $t(11) = 4.23$, $p < 0.01$. This may be related to the relative visibility of critical object-identifying features in different rotation angles. A one-way ANOVA showed no main effect of yaw angles on miss rate ($F(16, 187) = 1.50$, $p = 0.096$, $\eta^2 = 0.12$). With reference to the example in Figure 4, the views with negative pitch angle and large or zero yaw angles are further away from the canonical view (the viewpoint that maximizes the amount of information about an object; Palmer, 1981), which may contribute to the large miss rate of those viewpoints. One possible way to measure the difference of one view with the canonical view is a similarity metric in convolutional neural network (CNN) features, but it is out of the scope of the current study. While there is an effect of viewpoints angle on error rate, there is no effect on RTs (all $ps > 0.05$).

Figure 6 shows mean RTs on target-present trials as a function of visual set size, for each of the four memory set sizes. It is clear that the effects of visual set size were quite linear, as is typical in the search for one object among many (Vickery, King, & Jiang, 2005). A similar effect occurred for the target-absent trials (Figure 7) but with higher slopes.

We conducted a three-way, repeated-measures ANOVA on target-present RTs using condition (single vs. variable viewpoint), visual set size, and memory set size as factors. All three main effects were significant: condition, $F(1,11) = 29.56$, $p < 0.001$, $\eta_p^2 = 0.73$; visual set size, $F(2, 22) = 102.98$, $p < 0.001$, $\eta_p^2 = 0.90$; and memory set size, $F(3, 33) = 96.52$, $p < 0.001$, $\eta_p^2 = 0.90$. All the two-way interactions were significant: visual set size and condition, $F(2, 22) = 16.59$, $p = 0.001$, $\eta_p^2 = 0.60$; memory set size and condition, $F(3, 33) = 5.52$, $p < 0.01$, $\eta_p^2 = 0.33$; and visual set size and memory set size, $F(6, 66) = 50.39$, $p < 0.001$, $\eta_p^2 = 0.82$. Finally, the three-way interaction between visual set size, memory set size, and condition was also significant, $F(6, 66) = 4.13$, $p < 0.05$, $\eta_p^2 = 0.27$.

A similar pattern was seen in target-absent trials. However, there was no interaction between visual set size and condition, nor was there a three-way interaction between visual set size, memory set size, and condition. A three-way, repeated-measures ANOVA conducted on target-absent RTs revealed main effects of condition (single or variable), $F(1, 11) = 27.82$, $p < 0.001$, $\eta_p^2 = 0.72$; visual set size, $F(2, 22) = 92.78$, $p < 0.001$, $\eta_p^2 = 0.89$; memory set size, $F(3, 33) = 90.06$,

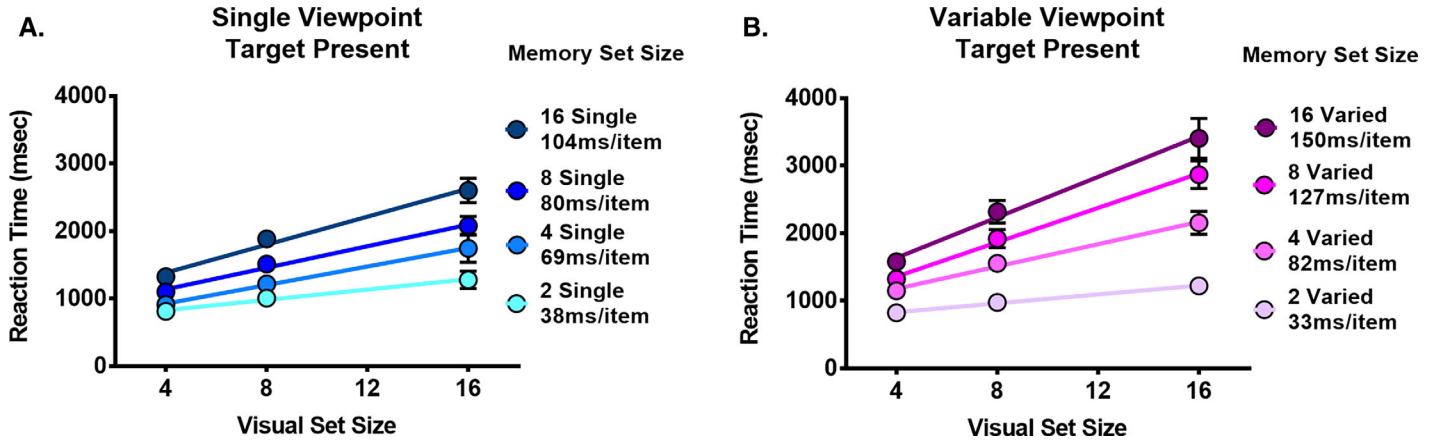


Figure 6. Reaction time on target-present trials as a function of visual set size in Experiment 2. (**A**) Data in single-viewpoints condition. (**B**) Data from the variable-viewpoint conditions (error bars: ± 1 *SEM*).

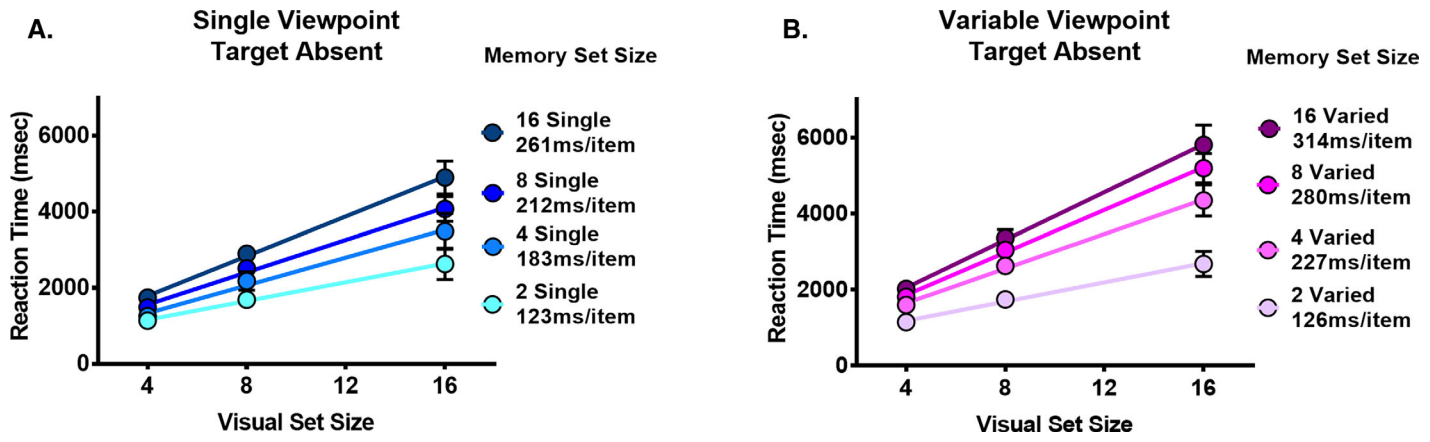


Figure 7. Reaction time on target-absent trials as a function of visual set size in Experiment 2. (**A**) Data in single-viewpoints condition. (**B**) Data from the variable-viewpoint conditions (error bars: ± 1 *SEM*).

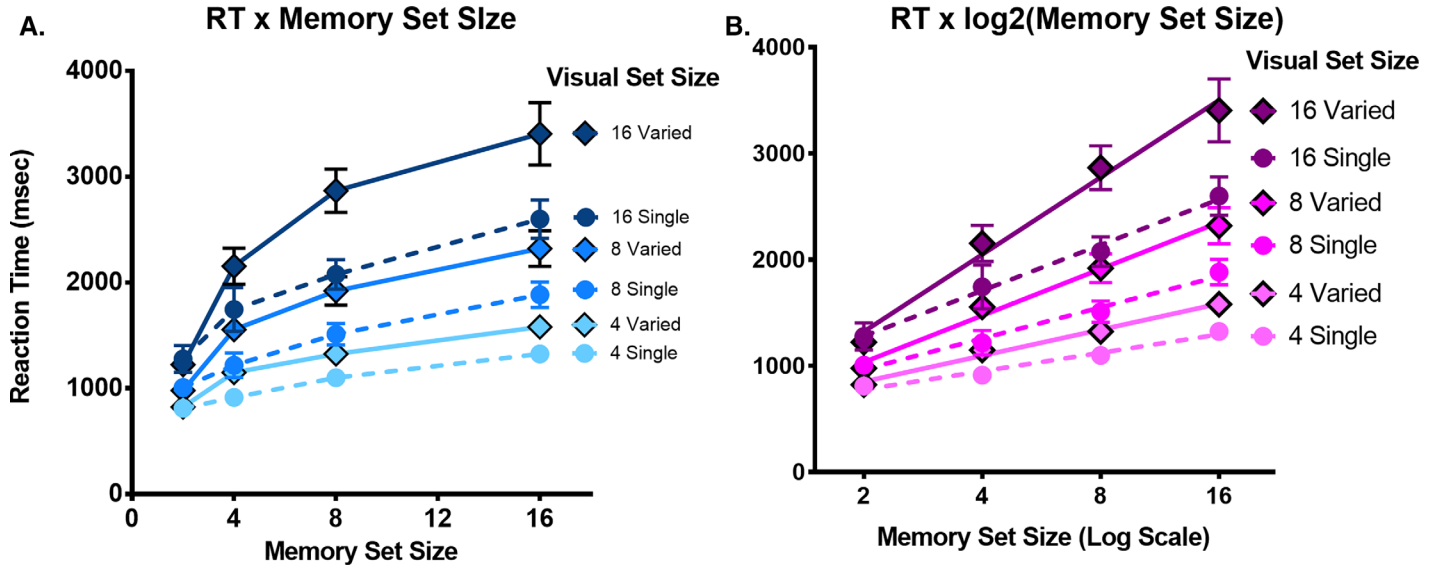


Figure 8. Reaction time on target-present trials as a function of memory set size. (**A**) Results on a linear x-axis. (**B**) Results on a logarithmic x-axis. Variable-viewpoint conditions are in solid lines; single-viewpoint conditions are plotted with dotted lines (error bars: ± 1 *SEM*).

$p < 0.001$, $\eta_p^2 = 0.89$; and significant interactions between visual set size and condition, $F(2, 22) = 20.61$, $p < 0.001$, $\eta_p^2 = 0.65$, and between visual set size and memory set size, $F(6, 66) = 44.55$, $p < 0.001$, $\eta_p^2 = 0.80$.

Figure 8A shows mean RTs as a function of memory set size. Note that the same RTs are plotted in Figures 6 and 8, in one case as a function of visual set size and in the other case as a function of memory set size. While the effect of visual set size on RT is basically linear in Figure 6, it is clear that this is not the case for the effect of memory set size where the functions in Figure 8A are curvilinear. Figure 8B replots the data on a logarithmic x-axis, showing that RT × log2(memory set size) functions are essentially linear. Linear regression on the RT × memory set size functions produces R^2 values of 0.75–0.98. Linear regression of RT × log2(memory set size) produces higher R^2 in all cases: 0.94–1.00. Error rates are larger at the larger set sizes, possibly reflecting some speed–accuracy trade-off, but overall, there is a convincingly linear relationship between RT and log2(memory set size).

Recall that the starting question for this experiment is whether searching for targets that can vary in their viewpoint is the same as searching for a single viewpoint. Figure 9 indicates that the single and variable conditions are similar for one or two targets. However, as the memory set size gets larger, search becomes less efficient. This indicates that the memory

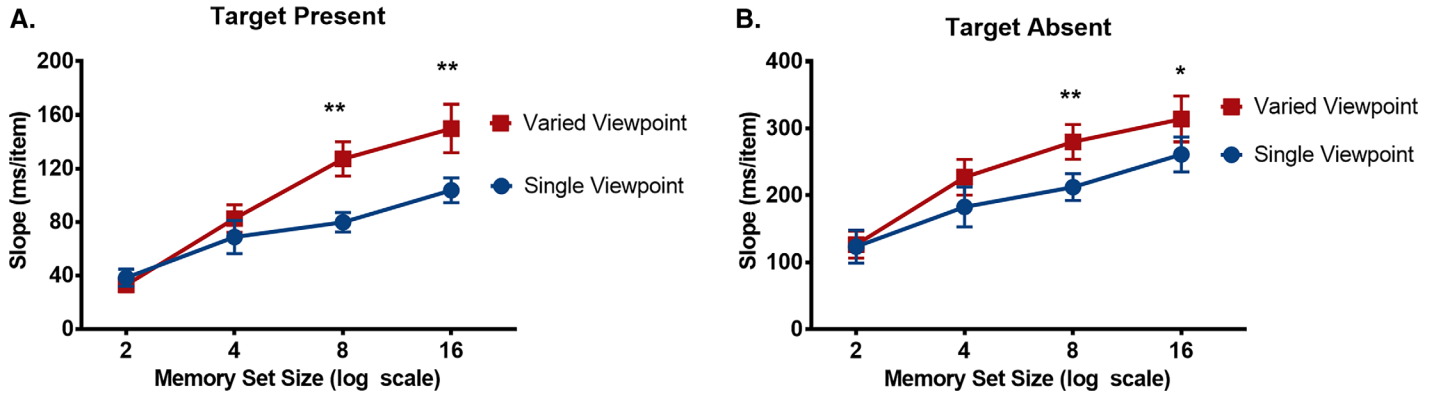


Figure 9. Slopes of RT × visual set size functions as a function of memory set size for single and variable-viewpoint conditions: (**A**) target present and (**B**) target absent (error bars: ± 1 *SEM*).

search, required each time a visual item is selected, is slower for variable viewpoints. One thing to note is that the lines in Figure 9 are connected lines. Each data point represents a slope value (y-axis) of RT × visual set size functions for four memory set sizes (x-axis). For Memory set size 2, there is no difference between conditions (variable: 33 ms/item; single: 38 ms/item; $t(11) = 0.88$, $p = 0.40$). There is a numerical difference for four targets (82 vs. 69 ms/item), but it is not statistically significant ($t(11) = 1.09$, $p = 0.30$). For the larger set sizes, the differences are significant (memory set size 8; 127 vs. 80 ms/item; $t(11) = 4.63$, $p < 0.001$; memory set size 16; 150 vs. 103 ms/item; $t(11) = 3.36$, $p < 0.01$). The variable slopes for larger memory set sizes resemble a categorical search slope (e.g., eight categorical targets: 125 ms/item; Cunningham & Wolfe, 2014). Similar results were found for the target-absent trials. This suggests that a small number of viewpoint-independent representations can be activated in hybrid search and can be checked without much more effort than the representation of a single view of a target. For larger memory sets, the viewpoint-independent representations take more time to examine. Experiment 3 examines two possible contributions to these results.

Experiment 3—enhancing the memory representation

Perhaps variable-viewpoint search was harder than single-viewpoint search because we did not adequately teach the Os about the different views of the targets in the variable-viewpoint condition. Experiment 3 is a variant of Experiment 2, intended to test that hypothesis by inducing the Os to more adequately encode a variable-viewpoint template or representation of the targets. In the variable-viewpoint condition of Experiment 2, observers only saw the canonical view of the targets during the initial memory session. In Experiment 3, subjects were exposed to all 119 viewpoints in the learning phase before starting the search trials. During the learning phase, only the targets from the memory sets were presented and rotated in 119 viewpoints, one viewpoint at a time. Each viewpoint was seen for approximately 80 ms, resulting in 10 s of total exposure to each target object. The number of memory test trials was still twice the memory set size for that block (50% "old," 50% "new"). A second change involves the viewpoints of the distractors. For the single-viewpoint condition in Experiment 2, all distractors were presented from the "canonical" view. In the variable-viewpoint condition, each distractor could be presented in any of the possible viewpoints. In Experiment 3, we equated the distractor sets in the two conditions by allowing the full range of viewpoints for distractors to be used in the single-viewpoint condition as well. In this experiment, for simplicity, only two memory set sizes (2, 8) and two visual set sizes (4, 16) were tested. Obviously, this does not allow us to test the shape of the RT × set size functions, but our interest here is specifically in the difference between single- and variable-viewpoint conditions. Eleven observers participated in this experiment. All other experimental conditions are the same as in Experiment 2.

Results

One participant whose average error rate was more than 20% was excluded from further analysis, leaving a total of 10 participants (mean age = 25.1 years, $SD = 8.0$, seven females). The average error rate was 7.9%, and these error trials were eliminated in the subsequent RT analyses. In addition, trials with RTs less than 200 ms or larger than 15,000 ms were excluded as outliers (0.01%). A paired sample t test on arcsine-transformed error rates revealed a higher miss

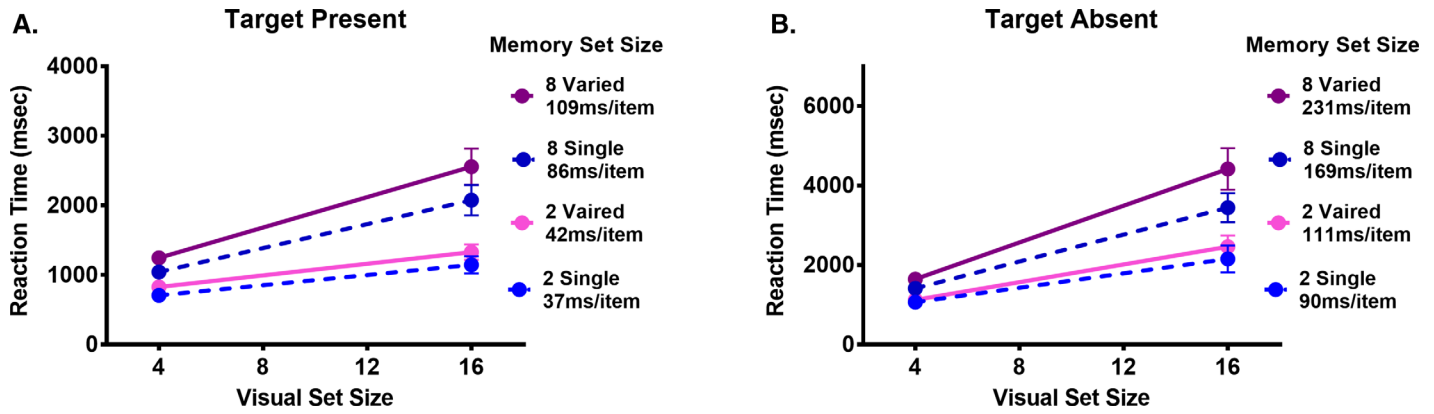


Figure 10. Reaction time on target-present and target-absent trials as a function of visual set size in Experiment 3. (**A**) Target present. (**B**) Target absent (error bars: ± 1 *SEM*).

rate for the variable-viewpoint condition ($M = 9.0\%$) than the single-viewpoint condition ($M = 6.9\%$), $t(9) = -5.27$, $p < 0.001$.

Figure 10 shows mean RTs on (a) target-present and (b) target-absent trials as a function of visual set size, for each of the two memory set sizes. Target-present and target-absent results are similar with higher slopes for the target-absent trials. We conducted a three-way, repeated-measures ANOVA on target-present RTs with condition (single vs. variable viewpoint), visual set size, and memory set size as factors. All three main effects were significant: condition, $F(1, 9) = 23.73$, $p = 0.001$, $\eta_p^2 = 0.73$; visual set size, $F(1, 9) = 44.29$, $p < 0.001$, $\eta_p^2 = 0.83$; and memory set size, $F(1, 9) = 42.76$, $p < 0.001$, $\eta_p^2 = 0.83$. Significant two-way interactions were found between visual set size and condition, $F(1, 9) = 6.87$, $p < 0.05$, $\eta_p^2 = 0.43$, and visual set size and memory set size, $F(1, 9) = 22.45$, $p = 0.001$, $\eta_p^2 = 0.71$. A similar pattern was seen in target-absent trials, with the addition of a significant two-way interaction between visual set size and condition and of a significant three-way interaction between visual set size, memory set size, and condition. Specifically, a three-way, repeated-measures ANOVA conducted on target-absent RTs revealed main effects of condition (single or variable), $F(1, 9) = 5.05$, $p = 0.05$, $\eta_p^2 = 0.36$; visual set size, $F(1, 9) = 50.62$, $p < 0.001$, $\eta_p^2 = 0.85$; memory set size, $F(1, 9) = 42.55$, $p < 0.001$, $\eta_p^2 = 0.83$; and significant interactions between visual set size and condition, $F(1, 9) = 5.27$, $p < 0.05$, $\eta_p^2 = 0.37$; between visual set size and memory set size, $F(1, 9) = 335.31$, $p < 0.001$, $\eta_p^2 = 0.80$; and between visual set size and condition, $F(1, 9) = 5.27$, $p < 0.05$, $\eta_p^2 = 0.37$. Finally, the three-way interaction between visual set size, memory set size, and condition was also significant, $F(1, 9) = 5.85$, $p < 0.05$, $\eta_p^2 = 0.39$.

When the memory set size is 2, searching for variable-viewpoint targets (42 ms/item) was again just as fast as searching for single-viewpoint targets (37 ms/item), $t(9) = 0.70$, $p = 0.50$. However, when more targets (8) are stored in memory, searching for variable-viewpoint targets (109 ms/item) was significantly less efficient than searching for specific viewpoints (86 ms/item), $t(9) = 2.71$, $p < 0.05$. A similar effect was found for the target-absent trials. Thus, these results replicate the findings of Experiment 2. There is a cost to the use of variable-viewpoint targets, and that cost is seen once more than a few targets are loaded into the memory set.

Experiment 4—all-in-one hybrid search

Experiments 2 and 3 show that variable-viewpoint stimuli produce less efficient hybrid search once more than one or two such stimuli are used as target items. Are these variable-viewpoint targets behaving like categories? From the vantage point of hybrid search, is searching for your cat in any of its possible poses the same as searching for the category of "cat"? In order to assess this more directly, Experiment 4 compares hybrid search for specific viewpoints, variable viewpoints, and categorical targets, using the same set of stimuli as in Experiments 2 and 3.

Methods

For simplicity, as in Experiment 3, only memory set sizes of 2 and 8 and visual set sizes of 4 and 16 were used. In the categorical search condition, we sampled target objects with replacement due to the limited number of exemplars per category (this number ranges from 10–43, as stated in Experiment 2). Previous

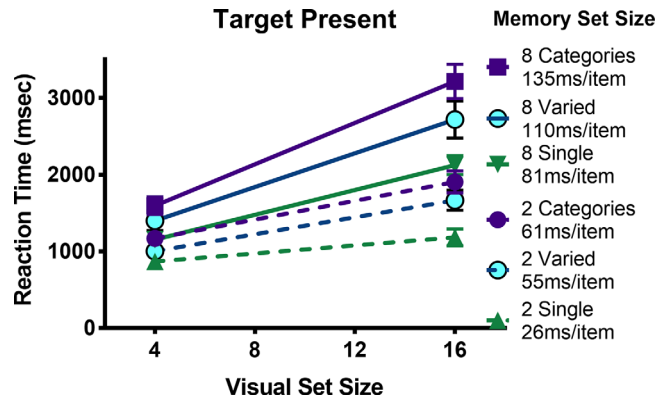


Figure 11. Reaction time on target-present trials as a function of visual set size in Experiment 4. Data in memory set size 8 condition are plotted with solid lines. Data from the memory set size 2 conditions are plotted with dotted lines (error bars: ± 1 *SEM*).

research on the role of familiarity of targets and distractors in hybrid search shows that such repetition makes little difference in the results (Wolfe, Boettcher, Josephs, Cunningham, & Drew, 2015). Similar RTs were obtained when the familiarity of targets and distractors was balanced and even in conditions where some distractors were made to be more familiar than the targets. Other experiment settings and apparatus are the same as in Experiment 2. The six blocks of Experiment 4 were run in pseudorandom order and counterbalanced across participants. Observers proceeded to the formal experiment after passing the memory test. Accuracy feedback was provided to observers after each search trial. Twelve observers (mean age = 24.1 years, $SD = 6.6$, eight females) participated in this experiment.

Results

One participant, whose average error rate was more than 20% in a condition, was excluded from further analysis, leaving a total of 11 participants. The average error rate was 8.5%, and those error trials were eliminated in the subsequent analysis. In addition, trials with RTs less than 200 ms or larger than 15,000 ms were excluded as outliers (removing just 0.1%). A paired sample t test on arcsine-transformed error rates revealed higher error rates in the categorical ($M = 9.8\%$) and variable-viewpoint conditions ($M = 9.4\%$) than in the single-viewpoint condition ($M = 6.2\%$, both ps < 0.05). There was no significant difference between the errors for the categorical and the variable-viewpoint conditions ($p = 0.53$).

The main results, shown in Figure 11, show performance with variable-viewpoint stimuli to lie between single-view and categorical stimuli. A three-way, repeated-measures ANOVA conducted on target-present RTs revealed main effects of condition (single, variable, or categorical), $F(2, 20) = 43.67$ $p < 0.001$, $\eta_p^2 = 0.81$; memory set size, $F(1, 10) = 71.78$, $p < 0.001$, $\eta_p^2 = 0.88$; visual set size, $F(1, 10) = 120.21$, $p < 0.001$, $\eta_p^2 = 0.92$; and significant interactions between visual set size and condition, $F(2, 20) = 19.11$, $p < 0.001$, $\eta_p^2 = 0.66$, and visual set size and memory set size, $F(1, 10) = 212.62$, $p < 0.001$, $\eta_p^2 = 0.96$. A similar pattern was found with the target-absent RTs. A three-way, repeated-measures ANOVA conducted on target-absent RTs revealed main effects of condition, $F(2, 20) = 51.64$ $p < 0.001$, $\eta_p^2 = 0.84$; memory set size, $F(1, 10) = 64.52$, $p < 0.001$, $\eta_p^2 = 0.87$; visual set size, $F(1, 10) = 112.51$, $p < 0.001$, $\eta_p^2 = 0.92$; and significant interactions between visual set size and condition, $F(2,$

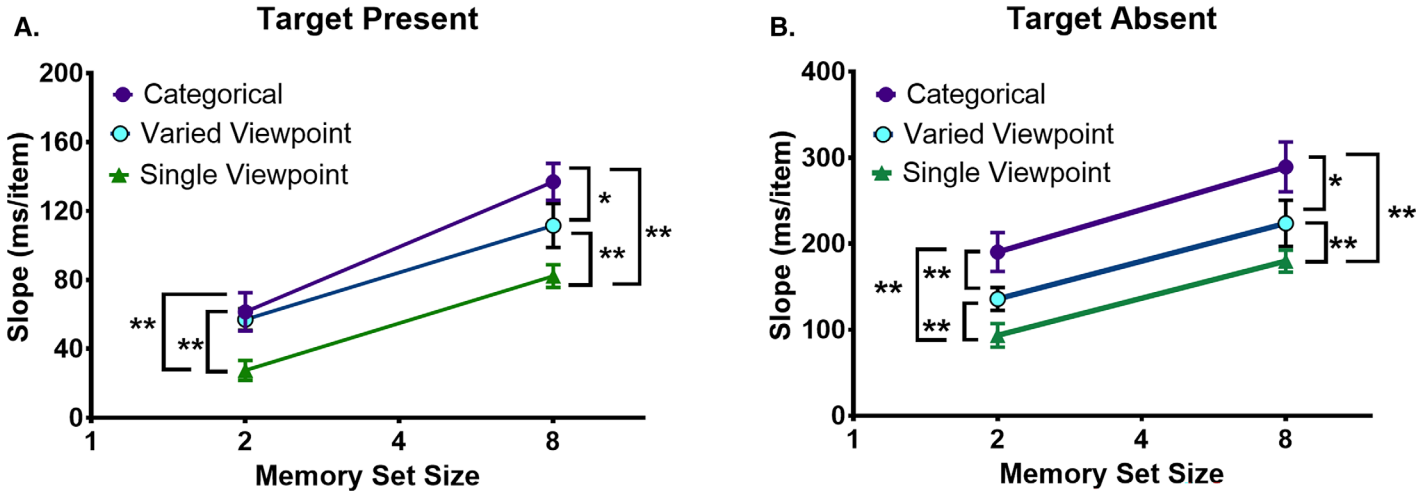


Figure 12. Slope of RT × visual set size functions for the different conditions of Experiment 4 at memory set sizes of 2 and 8. (**A**) Target present. (**B**) Target absent (error bars: ± 1 *SEM*).

20) = 27.07, $p < 0.001$, $\eta_p^2 = 0.73$, and visual set size and memory set size, $F(1, 10) = 78.46$, $p < 0.001$, $\eta_p^2 = 0.89$.

We further calculated RT × visual set size search slopes for each target condition and for each memory set size (Figure 12). We examined the interaction with paired sample *t* tests. Figure 12 shows the slopes of three conditions and two memory set sizes. Generally speaking, search for categorical targets in memory is less efficient than viewpoint variable targets, $t(10) = 2.09$, $p < 0.05$, and the specific viewpoint targets, $t(10) = 6.30$, $p < 0.001$, in large memory set size 8. For small memory set size 2, search for categorical targets in memory is still less efficient than the specific viewpoint targets, $t(10) = 3.70$, $p < 0.01$. But the search slope between the categorical targets and variable-viewpoint targets condition in small memory set size is statistically nonsignificant. For target-absent trials, the search slopes are shown to have significant differences among all three conditions (all *p*s < 0.05).

The findings suggest that the search difficulty is increasing from specific viewpoints to variable viewpoints and then to categorical targets. This seems reasonable if the time required to match a visual item to a memory set item increases with the abstractness of the memory set. Interestingly, one could imagine a different outcome where it would be easier to determine that an item in the world was, categorically, *a cat* than to determine that it was *my cat* or *my cat in a specific pose*. Perhaps we would see a different pattern of results in a version of the task where targets were more ambiguous. Regardless, the results suggest that we would see the standard hybrid search pattern of results in all cases.

Discussion and conclusions

In the end, this is a rather simple story. The core finding of hybrid search is that response times increase linearly with increasing visual set size, but they increase logarithmically with memory set size. If you want to know if any of your friends are present at a party, this logarithmic compression allows you to perform that hybrid search in a reasonable amount of time. At least that should be the case, as long as the hybrid search results are generally true and not simply limited to memory for specific pictures of specific objects in specific poses. The central finding of this article is that, when the stimuli sets were changed to scenes or to objects with variable viewpoints, we still replicated previous findings in hybrid search experiments. RTs remained linear functions of visual set size and were generally logarithmic functions of memory set size. The results did not need to come out this way. We presume that logarithmic RT × memory set size functions are logarithmic because, as Leite and Ratcliff (2010) propose, this is a consequence of trying to hold false-positive errors roughly constant. If either scenes or viewpoint variation jumbled the internal calculation of "set size," the results could have changed. In fact, whatever the underlying cause of the log functions, it is a very robust effect.

The effects of change in the hybrid task are seen in the overall difficulty of the different tasks. It was interesting to find that the search for photos of scenes was surprisingly difficult in Experiment 1. One might have predicted that scenes would support relatively fast search, given, for example, that the gist of a scene is extracted extremely fast (e.g., Thorpe, Fize, & Marlot, 1996). Why, then, is hybrid search for scenes less efficient than hybrid search for single objects? This may reflect the difference between determining general characteristics of a scene (indoor vs. outdoor, navigable vs. not navigable; Greene & Oliva, 2009) and determining that a visual scene matches a specific scene, stored in memory. Imagine a brief view of a scene. In a short amount of time, you could probably determine that this was an enclosed indoor room. You might recognize if it is a bedroom or an office. However, if asked whether a specific bookshelf or a computer was presented in the scene, you would need more time.

The role of clutter or crowding could be another factor explaining why hybrid search for objects is faster than hybrid search for scenes. An isolated single letter or object may be quite easy to identify, even if this item appears in the peripheral vision. However, if the same item was flanked by additional items in the periphery, it would become significantly harder to recognize (Rosenholtz, Huang, & Ehinger, 2012). The clutter of multiple complex scenes might slow search. Moreover, in object search, target items have unique outlines/shapes, whereas all scene images have rectangular outlines. The present data are suggestive, but testing these specific hypotheses would require further experiments in the future. Again, note, however, that these harder stimuli still produce standard hybrid results.

In the real world, objects will appear from various viewpoints. They will certainly not be limited to a canonical view. When we search for an object, we do not search for this exact image of an object; we search for this object, which could be viewed from various positions. The results of Experiments 2–4 tell us that hybrid search among different viewpoints is slower and less efficient than search for single-viewpoint targets even while those results replicate the logarithmic relationship of RT to memory set size. The difference between the efficiency of single- versus variable-viewpoint hybrid search is significant for large memory set sizes and less significant in small set sizes. Large set sizes serve to magnify effects. In this case, if every variable-viewpoint stimulus is a bit more difficult to process, then a large visual set size serves to

sum a collection of modest differences into one more substantial difference.

Hybrid search for variable viewpoints is faster and more efficient than categorical hybrid search, which suggests recognition of different viewpoints requires less time (or cognitive resources) than recognition of category membership. It would be interesting to determine if this is true in general or whether the level of category (e.g., "intermediate level," Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; or "entry level," Jolicoeur, Gluck, & Kosslyn, 1984) makes a difference.

Ideally, if we want to ask about hybrid search in the real world, it would be desirable to have observers search for a memorized set of objects among other objects in a real (or, at least, realistic) three-dimensional scene. If one imagines any reasonably interesting set of scenes, this is a daunting experiment to design because of the strong constraints that the scene would place upon the type, size, and viewpoint of the objects in the scene (Kallmayer, Võ, & Draschkow, 2023; Vo & Henderson, 2009). Still, it seems entirely likely, given the present results and previous work on hybrid search, that the basic hybrid search pattern would be seen with these hypothetical stimuli as well.

Keywords: hybrid visual and memory search, scene, viewpoint, category

Acknowledgments

Supported in part by the National Institutes of Health (EY017001) to JMW and National Natural Science Foundation of China (62206015, 62227801, U20B2062) to BZ. We thank K. Ehinger for her valuable suggestions. This work was done while the first three authors were in Visual Attention Lab, Harvard Medical School and Brigham & Women's Hospital, United States.

Commercial relationships: none.
Corresponding author: Bochao Zou.
Email: zoubochao@ustb.edu.cn.
Address: School of Computer and Communication Engineering, University of Science and Technology, Beijing, China.

References

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bulthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1506–1514.

Boettcher, S. E. P., & Wolfe, J. M. (2015). Searching for the right word: Hybrid visual and memory search for words. *Attention, Perception, & Psychophysics, 77*(4), 1132–1142.

Brainard, H. D. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436, https://doi.org/10.1007/s13398-014-0173-7.2.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., . . . Yu, F. (2015). *ShapeNet: An Information-Rich 3D Model Repository.* Issue arXiv:1512.03012 [cs.GR].

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.

Cunningham, C. A., & Wolfe, J. M. (2014). The role of object categories in hybrid visual and memory search. *Journal of Experimental Psychology. General, 143*(4), 1585–1599, https://doi.org/10.1037/a0036313.

Drew, T., Boettcher, S. E. P., & Wolfe, J. M. (2016). Searching while loaded: Visual working memory does not interfere with hybrid search efficiency but hybrid search uses working memory capacity. *Psychonomic Bulletin & Review, 23*(1), 201–212.

Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: Computer applications* (pp. 231–243). Dordrecht: Springer Netherlands.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20*(4), 464–472.

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology, 16*(2), 243–275.

Kallmayer, A., Võ, M. L.-H., & Draschkow, D. (2023). Viewpoint dependence and scene context effects generalize to depth rotated three-dimensional objects. *Journal of Vision, 23*(10), 9, https://doi.org/10.1167/jov.23.10.9.

Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics, 72*(1), 246–273.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Palmer, S. E. (1981). Canonical perspective and the perception of objects. *Attention and Performance, 9*, 135–151.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382–439.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology, 3*, 13.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: II. Detection, search, and attention. *Psychological Review, 84*(1), 1.

Tarr, M. J., & Bulthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1494–1505.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520.

Vickery, T. J., King, L.-W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision, 5*(1), 8, https://doi.org/10.1167/5.1.8.

Vo, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3), 1–15, https://doi.org/10.1167/9.3.24.

Wolfe, J. M. (2012). Saved by a log: How do humans perform hybrid visual and memory search? *Psychological Science, 23*(7), 698–703, https://doi.org/10.1177/0956797612443968.

Wolfe, J. M. (2020). Visual search: How do we find what we are looking for? *Annual Review of Vision Science, 6*, 539–562, https://doi.org/10.1146/annurev-vision-091718-015048.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, and Psychophysics, 73*(6), 1650–1671, https://doi.org/10.3758/s13414-011-0153-3.

Wolfe, J. M., Boettcher, S. E. P., Josephs, E. L., Cunningham, C. A., & Drew, T. (2015). You look familiar, but I don't care: Lure rejection in familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1576–1587, https://doi.org/10.1037/xhp0000096.

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review, 28*(4), 1060–1092.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492). San Francisco, CA, USA: IEEE.