Cube search, revisited

Xuetao Zhang
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China

Jie Huang
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Serap Yigit-Elliott
Exponent, Bellevue, WA, USA

Ruth Rosenholtz
Department of Brain and Cognitive Sciences, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

Observers can quickly search among shaded cubes for one lit from a unique direction. However, replace the cubes with similar 2-D patterns that do not appear to have a 3-D shape, and search difficulty increases. These results have challenged models of visual search and attention. We demonstrate that cube search displays differ from those with "equivalent" 2-D search items in terms of the informativeness of fairly low-level image statistics. This informativeness predicts peripheral discriminability of target-present from target-absent patches, which in turn predicts visual search performance, across a wide range of conditions. Comparing model performance on a number of classic search tasks, cube search does not appear unexpectedly easy. Easy cube search, per se, does not provide evidence for preattentive computation of 3-D scene properties. However, search asymmetries derived from rotating and/or flipping the cube search displays cannot be explained by the information in our current set of image statistics. This may merely suggest a need to modify the model's set of 2-D image statistics. Alternatively, it may be *difficult* cube search that provides evidence for preattentive computation of 3-D scene properties. By attributing 2-D luminance variations to a shaded 3-D shape, 3-D scene understanding may slow search for 2-D features of the target.

Vision is not the same everywhere

In visual search, an observer looks for a particular *target* item among other items known as *distractors*. Intriguingly, search is sometimes difficult even when an observer can clearly distinguish the target from the distractors. For example, search for a randomly oriented T among randomly oriented Ls is difficult (Wolfe, Cave, & Franzel, 1989), even though we can easily tell an individual T from an L. Similarly, search for a target defined by a conjunction of features—such as a white vertical bar among black verticals and white horizontals—is difficult relative to a feature search for a horizontal bar among verticals or for a white bar among black (Treisman & Gelade, 1980; Treisman & Schmidt, 1982). These phenomena imply that vision is not the same everywhere. If it were, the easy discriminability of focal target and distractor pair should lead to easy search.

In what way is vision not the same everywhere? Popular models of search have focused on potential differences between attended and unattended vision. This not only includes theories such as the seminal Feature Integration Theory (Treisman & Gelade, 1980) and later Guided Search (Wolfe, 1994) but is also at least implicit in many other theories of search (e.g., Itti, Koch, & Niebur, 1998; Li, 2002; Rosenholtz, 1999; Torralba, Oliva, Castelhano, & Henderson, 2006; Zhang, Tong, Marks, Shan, & Cottrell, 2008). However, a number of researchers have shown results that are inconsistent with the notion that differences between attended and unattended vision drive search performance (e.g., Carrasco, Evert, Chang, & Katz, 1995; Carrasco, McLean, Katz, & Frieder, 1998; Reddy & VanRullen, 2007; Vlaskamp, Over, & Hooge, 2005;

Citation: Zhang, X., Huang, J., Yigit-Elliott, S., & Rosenholtz, R. (2015). Cube search, revisited. *Journal of Vision, 15*(3):9, 1–18, http://www.journalofvision.org/content/15/3/9, doi:10.1167/15.3.9.

doi: 10.1167/15.3.9 Received July 8, 2014; published March 16, 2015 ISSN 1534-7362

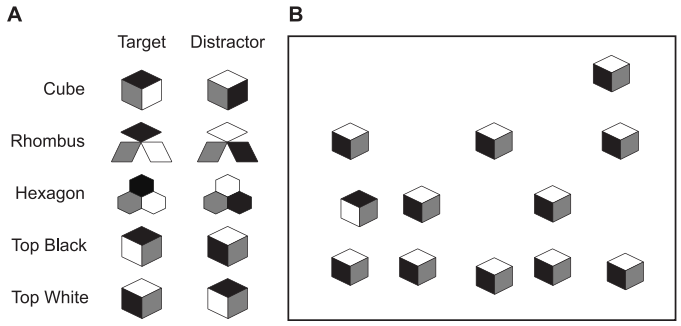

Figure 1. Stimuli. (A) Target–distractor pairs, modeled after Enns and Rensink (1990a). (B) An example 12-item layout from Enns and Rensink's (1990a) experiments (Top Black condition).

Wertheim, Hooge, Krikke, & Johnson, 2006; Wolfe, 1994) or have offered alternative explanations (e.g., Eckstein, 1998; Geisler & Chou, 1995; Gheri, Morgan, & Solomon, 2007; Palmer, Ames, & Lindsey, 1993; Palmer, Verghese, & Pavel, 2000; Rosenholtz, Huang, & Ehinger, 2012; Verghese & Nakayama, 1994). One puzzle is the ease of search for 3-D scene properties such as lighting direction and 3-D orientation, as discussed later.

Recently, researchers have suggested an alternative view of search in which the key way that vision is not the same everywhere is instead due to the differences between foveal and peripheral vision (Carrasco et al., 1995; Carrasco & Frieder, 1997; Carrasco et al., 1998; Carrasco & Yeshurun, 1998; Geisler & Chou, 1995; Gheri et al., 2007; Najemnik & Geisler, 2005, 2008, 2009; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Vlaskamp et al., 2005; Wertheim et al., 2006). In this article, we will examine a few search results which were problematic for the attention story, in light of known limits of peripheral vision.

Attention-based models and search for 3-D scene properties

In the standard attention story, visual processing proceeds from low-level through higher level mechanisms. Up to some point in the pipeline, processing occurs in parallel across the visual field. However, capacity limits in vision prohibit higher level processing from occurring everywhere at once. At some stage, attention selects only a portion of the visual input for later processing.

According to this classic view, search difficulty allows us to pinpoint at what stage selective attention operates. If attentional selection operates late, after all items in the display have been processed to the point of identification, then search should be easy so long as the target and distractors are visually distinct. If attentional selection operates before, say, shape-from-shading computations, then search should be difficult when target and distractors contain similar 2-D features but differ in their apparent 3-D shape. Following this logic, early search results suggested that selective attention operates just after the computation of spatially organized maps of basic features. Conjunction or configuration of these basic features requires the serial deployment of selective attention to "bind" features together. This model is known as Feature Integration Theory (Treisman & Gelade, 1980).

The explanatory power of Feature Integration Theory hinges on specifying the set of preattentive basic features. Researchers have in general agreed upon a few, seemingly relatively low-level basic features, such as orientation, color, size, and motion (Wolfe & Horowitz, 2004). However, other search experiments have suggested that higher level features may also be available preattentively.

For example, Enns and Rensink (1990a) found that search for a "bottom-lit" cube among top-lit ones (Figure 1) was efficient (target-present slopes of 8 ms per item). To test whether this search task (Cube condition) was easy simply due to low-level cues, they also tested several 2-D "equivalents." In the first control (Rhombus condition), they rearranged the white, gray, and black polygons forming the cubes into abstract patterns that did not give rise to a 3-D percept. These stimuli maintain—loosely speaking—the arrangement and shape of the polygons comprising the cubes but not the junction between them. In the second control (Hexagon condition), Enns and Rensink preserved the Y-junction between the three faces, and their gray values, but not the shape of each face. Both the Rhombus and Hexagon conditions (Figure 1A) led to inefficient search (target-present slopes of 19 and 20 ms per item, respectively).

Observers could efficiently search for a black-topped polygonal pattern among white-topped patterns, but seemingly only when that pattern was interpretable as 3-D. This suggested that observers might use as a cue a higher level property of 3-D scenes, such as lighting direction. If so, performance might depend upon the ecological validity of the lighting direction (see also Ramachandran, 1988; Sun & Perona, 1996). In fact, Enns and Rensink found that search was easier when a cube target appeared bottom lit—and the distractors top lit—compared to when it was shaded as if under a more ecologically familiar lighting from above (see Figure 1A; Top Black target: 6 ms per item; Top White target: 21 ms per item). When the targets and distractors were turned upside down, the asymmetry reversed. The target with the black bottom (common lighting direction) yielded inefficient target-present search slopes (18 ms per item). The target with a white bottom (uncommon lighting) led to more efficient

search (10 ms per item). Again, search for a cube lit from an uncommon lighting direction, among distractors lit from a common lighting direction, appears to be easier. This parallels other results showing easier search for a less familiar target among familiar distractors than vice versa (Wang, Cavanagh, & Green, 1994).

Do these search results imply that lighting direction is a basic feature, available preattentively, like orientation or color? Ostrovsky, Cavanagh, and Sinha (2005) questioned this account, showing that search for a unique lighting direction is quite difficult when the cubes have random 3-D orientation or when the inconsistently lit target appears within a complex scene. Similarly, Rensink and Cavanagh (2004) showed that search for a tilted quadrilateral becomes less efficient when one *adds* a cue that the target is lit from a unique direction (though search in this condition is still fairly efficient at 13–14 ms per item).

Nonetheless, the puzzle remains as to why Enns and Rensink's (1990a) cube search was more efficient than the 2-D equivalent conditions, and why flipping the displays upside down causes the search asymmetry to reverse. Enns and Rensink's results, as well of those of Sun and Perona (1996) and Ramachandran (1988), seem to suggest that some property of the 3-D scene, such as lighting direction, is available preattentively and in parallel across the visual field, supporting efficient search. Other properties of 3-D scenes also seem as if they might be available preattentively, including the 3-D orientation of a rectangular cuboid (Enns & Rensink, 1991).

These results pose a challenge for any model of visual search. Enns and Rensink (1990a, 1990b) suggested that early vision might have access to "spatial and intensity relations that convey three-dimensionality" (Enns & Rensink, 1990a, p. 722) and lighting direction, but noted that neither junctions nor intensity relations alone seemed sufficient; rather, easy search appeared to require a consistent percept of 3-D shape. This conclusion is problematic if selective attention operates at a single level of the visual processing hierarchy, as originally proposed: These 3-D scene properties are seemingly higher level than typical basic features such as orientation and color. (Unless, of course, search operates over a representation of orientation and color that is also produced relatively late, or in a side channel separate from the main processing pipeline, as in Wolfe's [1994] Guided Search.) Perhaps 3-D scene properties are important enough that the visual system has developed efficient, preattentive processing of those properties. Or perhaps selective attention operates based on information from multiple levels of the processing hierarchy (Allport, 1993; Reddy & VanRullen, 2007; Tsotsos et al, 1995; VanRullen, Reddy, & Koch, 2004; Wolfe & Horowitz, 2004) or is flexible as to at what level it operates (Di Lollo, Kawahara, Zuvic, & Visser, 2001; Nakayama, 1990; Treisman, 2006). These latter models, certainly, are so flexible that they are difficult to disprove.

More problematically, how could 3-D shape, reflectance, or lighting direction even be computed without the sorts of conjunctions and configurations of features supposedly unavailable without attention? How could it be that processing of 3-D shape, lighting, and/or reflectance occurs preattentively but simple feature binding does not? Does the visual system compute the necessary conjunctions to extract 3-D properties, only to throw away those conjunctions and recompute them when attention is present? This is not out of the question, given a sufficiently restrictive bottleneck. However, while there is little disagreement that the brain has limited capacity, it is not clear that visual cortex contains this degree of bottleneck.

Here we consider an alternative explanation for what makes search easy or difficult, to see if we can shed light on these puzzles. Enns and Rensink (1990a, 1990b) made a good attempt to test "equivalent" 2-D stimuli while having minimal knowledge of the relevant 2-D features. We reexamine their results in light of recent understanding of low-level features of relevance to peripheral vision and ask whether one feature set—albeit a complex one—can predict a range of search results.

Peripheral vision and search

An important way in which vision is not the same everywhere concerns the difference between the fovea and the periphery. Clearly peripheral vision is important in visual search. The periphery, being much larger than the fovea, is inherently more likely to contain the target, and typically the target is peripheral until it is found. Therefore, to understand search we must understand the strengths and limitations of peripheral vision (Carrasco & Frieder, 1997; Carrasco et al., 1998; Erkelens & Hooge, 1996; Geisler, Perry, & Najemnik, 2006; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj et al., 2012).

Peripheral vision has considerable loss of information relative to the fovea. This loss of information begins as early as the retina, which employs variable spatial resolution to get past the bottleneck of the optic nerve. However, the loss relative to the fovea does not end there; the reduced peripheral acuity has a modest effect when compared with *visual crowding*. An example of this phenomenon appears in the top of Figure 2. A reader fixating the central cross will likely have no difficulty identifying the isolated letter on the left. However, the same letter can be difficult to recognize when flanked by additional letters. Move the flankers farther from the target, and at some *critical spacing*, recognition is restored. Behavioral work suggests that

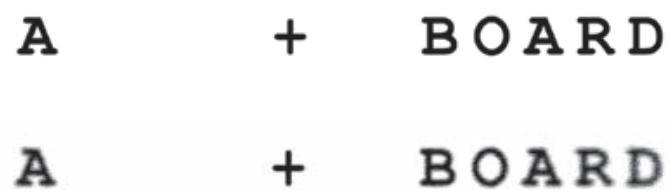

Figure 2. Visual crowding. Top: When fixating on the +, the *A* on the left is easy to recognize, whereas the *A* in the word "BOARD" can be quite difficult to identify. Bottom: This cannot be explained by a mere loss of acuity in peripheral vision. Here, we have mimicked loss of acuity using a Photoshop radial blur (spin = 1, zoom = 4), à la Anstis (1998). Note that this blur exaggerates the peripheral loss of acuity. Even with that loss of acuity, the crowded *A* on the right is quite readable.

the critical spacing is approximately 0.4 to 0.5 times the eccentricity (the distance to the center of fixation) for a fairly wide range of stimuli (Bouma, 1970; Pelli, Palomares, & Majaj, 2004).

These effects cannot be explained by lower acuity in the periphery (Lettvin, 1976). One can demonstrate this by filtering to mimic the loss of resolution (bottom of Figure 2). If lower acuity were responsible for difficulty reading the *A* in "BOARD," then when we look at that *A* in the transformed image, it should be difficult to read. Clearly the quite modest reduction in peripheral acuity does not predict difficulty reading the crowded letters. In fact, the blur shown here exaggerates the acuity loss; as Anstis did (1998), we have amplified the blur to make it more noticeable. The extent of the blur shown here is approximately 5 times that needed to mimic the loss of resolution at the eccentricity of the *A*s.

The occurrence of crowding when flankers lie within some critical spacing makes clear that peripheral vision processes sizable patches. In an often-cluttered search display, these patches likely contain multiple elements. The information in these patches may be useful for guiding search (which under normal conditions means guiding eye movements) or not. Consider Figure 3A. For a given fixation, one patch may contain the target plus a number of distractors. Another patch may contain only distractors. If it is easy for peripheral vision to distinguish between these two kinds of patches, search should be easy, as the information will immediately guide the observer to the target—or even simply allow the observer to identify that target peripherally. At the other extreme, if peripheral information cannot discriminate between target-present and target-absent patches, the observer will need to scan the display, looking for the target. In between, peripheral information guides search with intermediate amounts of reliability.

In previous work (Rosenholtz, Huang, Raj et al., 2012), we measured peripheral discriminability (d') and compared that to search difficulty. Subjects were asked whether the middle item in a crowded array was a target or distractor. The flankers were all distractors. We found that there was a good relationship between peripheral d' and search performance for five classic search conditions. This supports the hypothesis that search performance is constrained by the abilities and limitations of peripheral vision. It suggests that in order to predict the difficulty of a search task, one can measure the discriminability of target-present from target-absent patches in the periphery.

The Texture Tiling Model of peripheral vision

We can take this theory further by developing a model of the information lost and maintained in

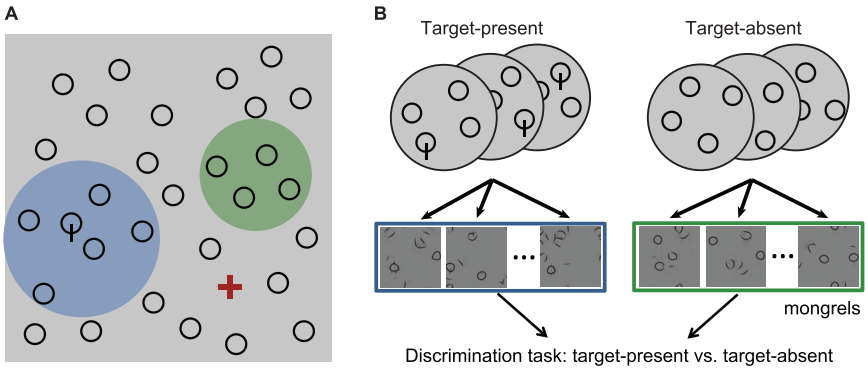

Figure 3. (A) A fixation (red cross) during visual search. Crowding indicates that the visual system processes sizable patches in the periphery; some contain both a target and distractors (blue), whereas most contain only distractors (green). The visual system needs to distinguish between promising and unpromising peripheral patches and move the eyes accordingly. (B) We hypothesize that peripheral patch discriminability critically limits search. Furthermore, a key limit on peripheral vision is the information available in a rich set of 2-D image statistics. To test the latter, we select a number of target-present and target-absent patches and use texture synthesis routines to generate a number of patches which have the same statistics (mongrels) but are otherwise random. We then ask human observers to discriminate between target-present and target-absent mongrels, and examine whether this discriminability predicts peripheral performance and search difficulty. Adapted from Rosenholtz et al., 2012b.

peripheral vision and relating performance to low-level mechanisms. Recent research has suggested that visual crowding is due to a texture-like representation in peripheral vision (Balas, Nakano, & Rosenholtz, 2009; Lettvin, 1976; Levi, 2008; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Pelli and Tillman, 2008). We have proposed (Balas et al., 2009) that rich, texture-like statistics are measured in parallel across the visual field within pooling regions that grow linearly with eccentricity, in accord with Bouma's law; we call this the Texture Tiling Model. To test this model, one needs a candidate set of statistics that are consistent with texture perception. A number of such candidates have been proposed (for a review, see Rosenholtz, 2014). The most successful set at present (as judged by ability to produce subjectively similar textures) is that identified by Portilla and Simoncelli (2000), and we adopt these statistics in our modeling. They are marginal distribution of luminance; luminance auto-correlation; correlations of the magnitude of responses of oriented V1-like wavelets across differences in orientation, neighboring positions, and scale; and phase correlation across scale. This is simpler than it may sound; computing a given second-order correlation merely requires taking responses of a pair of V1-like filters, point-wise multiplying them, and taking the average over the pooling region. Such second-order measurements are important for reducing redundancy in coding of natural images (Balas, 2006; Zetzsche, Barth, & Wegmann, 1993).

These image statistics have been shown to do a good job of capturing texture appearance (Portilla & Simoncelli, 2000). Mounting evidence suggests that such 2-D image statistics may underlie peripheral encoding, which is of great relevance for visual search: The informativeness of those same 2-D image statistics predicts performance at both peripheral recognition tasks (Balas et al., 2009; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj et al., 2012) and classic search tasks (Rosenholtz, Huang, Raj et al., 2012), and equating those local statistics creates visual metamers for natural scenes (Freeman & Simoncelli, 2011). Any successful parametric texture model can provide candidate texture statistics, but at present the best is that of Portilla and Simoncelli.

A rich, high-dimensional set of local image statistics provides an efficient, compressed representation (Rosenholtz, 2011), which captures a great deal of information about the visual input. Nonetheless, the encoding is lossy, meaning one cannot generally reconstruct the original image. We hypothesize that the information maintained and lost by this representation provides an important constraint on peripheral processing. In particular, if the available information cannot discriminate between target-present and target-absent patches, search will be inefficient.

How can we test this model? The holy grail of models is to go directly from measurements of the input to testable predictions. For instance, a number of researchers have developed models that can predict from image input the visibility of low-contrast patterns (see Rohaly, Ahumada, & Watson, 1997, for a review). Developing these models required measuring visibility thresholds for a number of sine-wave gratings. In practice, many successful models (e.g., Watson, 1993) have also measured how those thresholds vary as a function of the other spatial frequencies present in the display, i.e., with contrast masking. The models transform this handful of measured visibility thresholds for simple gratings into predictions of the visibility of a much wider array of stimuli.

We could similarly attempt to ground our model in a set of basic psychophysical measurements. We could measure discriminability thresholds for each of the image statistics, and by analogy with contrast masking we could measure how each threshold varies as a function of the other image statistics. Based on these threshold measurements, we could input an arbitrary pair of image patches and make predictions about their peripheral discriminability. However, given the number of image statistics in the Texture Tiling Model, this methodology would require the measurement of a very large number of thresholds—there are approximately 1,000 image statistics measured for each patch—and we do not currently have those measurements.

Another common methodology for testing models, when an end-to-end system is unavailable, involves putting a human somewhere "in the loop." For example, many perceptual grouping models process the stimulus and output an image showing candidate groups for a human to look at and evaluate. Other models rely on the experimenter to provide labeled features from the display. Here we put the human in the loop using more formal methods.

We hinted at the methodology in our discussion of Figure 2. We filtered the stimulus to model loss of high spatial frequencies in peripheral vision. By viewing the results in the fovea, where minimal additional loss of high-frequency information occurs, we observed that reduced peripheral acuity could not account for difficulty identifying letters in the crowded display. To be more rigorous, we could have let observers examine a number of images filtered in this way and measured how well they could identify the letters in each array. For stimuli like that in Figure 2, such an experiment would be a waste of time, as it is hard to imagine the observers would make any errors. Actual peripheral performance would degrade as one added nearby flankers, but performance when free-viewing the filtered images would stay flat at around 100% correct. In other words, the modeled reduction in acuity would not predict peripheral performance; such results would

demonstrate more formally that peripheral acuity loss cannot explain crowding.

Similarly, our model postulates the information lost and maintained by early peripheral processing. We can impose that information loss on the stimuli. We first blur an input patch to approximate the reduction in peripheral acuity, and we then measure image statistics on the blurred patch. We then start from a random noise image and use the Portilla and Simoncelli (2000) texture synthesis algorithm to iteratively apply constraints derived from each of the measured image statistics. By doing so, we generate "mongrels," i.e., images that have been synthesized to have approximately the same image statistics as the blurred patch but are otherwise random. We then experimentally ask observers to discriminate between mongrels originating from target-present versus target-absent patches (see Figure 3B). This provides us with a measure of the inherent discriminability of target-present from target-absent patches, given the hypothesized loss of peripheral information.

Geisler and Chou (1995) elucidated requirements for assessing the role of low-level factors in complex tasks, using a methodology related to that described here. Their methodology, like ours, compares human performance at a discrimination task to human performance on the complex task of interest. Our methodology follows the spirit of their procedure but differs, for example, in using a model to generate stimuli for the discrimination experiment, as described earlier. Geisler and Chou point out that in order to measure the impact of low-level factors with the discrimination task, we must minimize and normalize higher level factors. To this end, we use well-trained observers, who receive feedback throughout the experiment, and allow them to freely attend to and view the mongrels.

Free-viewing mongrels also protects us from the logical error of testing whether peripheral vision plus our model of peripheral vision can predict peripheral vision. Note that viewing the blurry "BOARD" in Figure 2 while fixating the + would have given the *wrong* answer; letter recognition would have been difficult, *incorrectly* suggesting that acuity loss provides a viable explanation for crowding.

If peripheral crowding is a result of our proposed low-level mechanism, discriminability of target-present mongrels from target-absent mongrels (mongrel d') should predict peripheral discriminability of target-present from target-absent patches (peripheral d'). In addition, if search is limited by information available in peripheral vision, mongrel d' should also predict search difficulty.

In the past, we have compared mongrel d', peripheral d', and search performance for classic search conditions. We found good relationships among all three performance measures (Rosenholtz, Huang, Raj, et al., 2012). This suggests both that our model of low-level parallel processing stages is predictive of peripheral performance under conditions of crowding and that the low-level loss of peripheral information acts as a critical factor in visual search.

Here we will apply the same methodology to reexamine cube search. We ask whether there is something special about cube search, once we factor out the low-level information available for the task. In essence, our model allows us to quantitatively compare the 2-D information available in the cube search displays to that in equivalent 2-D rhombus and hexagon displays. Is cube search easier than we would expect, based upon the difficulty of other, 2-D search tasks? If so, this would imply that cube search could make use of additional preattentive information, such as 3-D scene properties. To get at this question, we replicate five search tasks (Figure 1) from Enns and Rensink (1990a), using the same layout and experimental methodology we previously used for five classic search conditions (Rosenholtz, Huang, Raj et al., 2012). For all 10 search conditions, we will compare search results, peripheral discriminability, and mongrel discriminability. To the extent that peripheral discriminability provides a crucial determinant of search performance, it should predict search performance across the range of tasks. To the extent that our relatively low-level model predicts peripheral discriminability, mongrel d' should predict peripheral d' independently of whether the stimuli are 2-D or 3-D. It should appear that one function relates mongrel d' to peripheral d' for all conditions. By extension, the same should be true of predicting search from mongrel d'. On the other hand, if cube search is special, then when we plot the data, the 3-D tasks should be outliers relative to the 2-D tasks; the relationship for the 2-D tasks would not predict the 3-D tasks.

Predicting search for cubes and 2-D equivalents

Methods

Visual search task

Subjects: Ten subjects participated in the search experiment after giving written informed consent. The data from four of those subjects were excluded from further analysis, due to high error rates (>25% errors). All subjects had normal or corrected-to-normal vision and received monetary compensation for their participation.

Stimuli and procedure: To compare with our previous search results (Rosenholtz, Huang, Raj et al., 2012), we

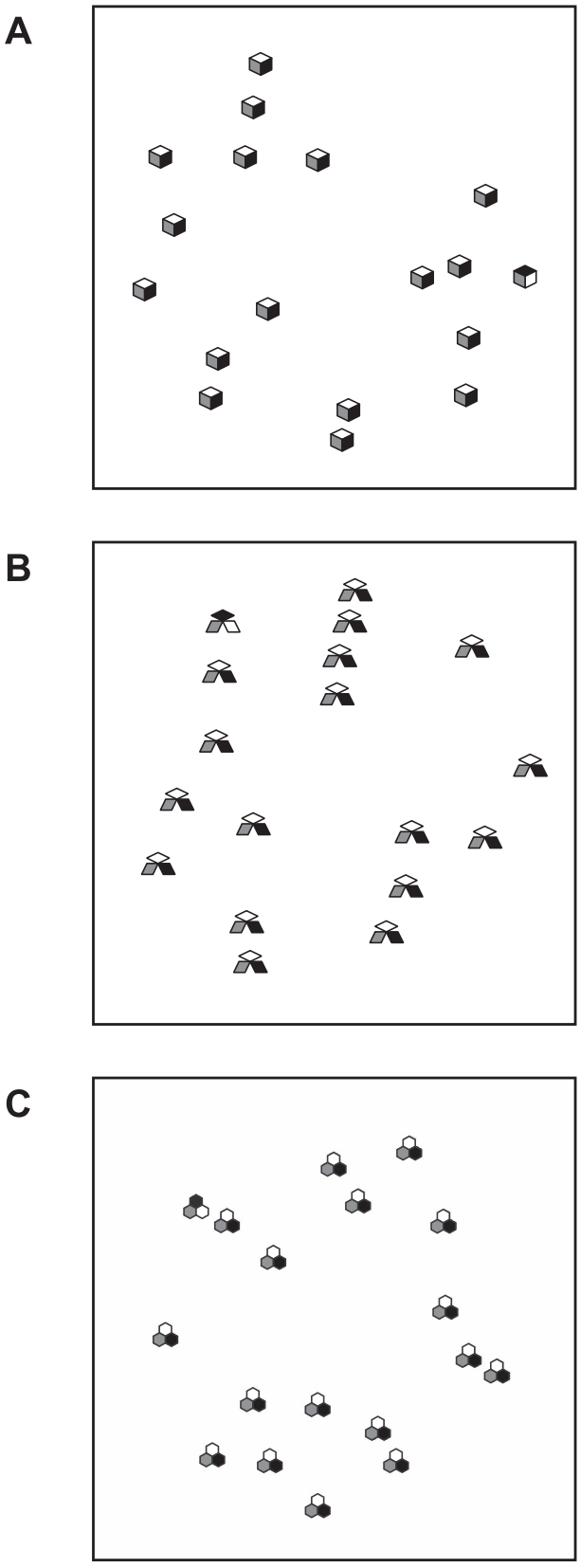

Figure 4. Sample target-present stimuli. (A) Cube condition. (B) Rhombus condition. The difficulty of these search tasks may not be scale invariant, so the displays should be viewed as large as possible to mimic experimental conditions. (C) Sample target-present stimulus for the Hexagon condition.

reran several of Enns and Rensink's (1990a) search experiments with the same item layout and methodology. Comparing Figure 1B and Figure 4 elucidates the difference in layout between Enns and Rensink's experiments and those reported here. An additional reason to normalize item layout across conditions is that we have previously found an easy cube search task (Sun & Perona, 1996) to be considerably less efficient using our item layout (Rosenholtz, Huang, & Ehinger, 2012), perhaps due to emergent features in the Sun and Perona displays. We did our best to replicate the appearance of the search items. Our search conditions differed from Enns and Rensink's in that our subjects were allowed to move their eyes during search.

We tested search for a bottom-lit cube among top-lit ones (Enns and Rensink's experiment 2A, our Cube condition) as well as search displays using some of Enns and Rensink's equivalent 2-D targets and distractors (their experiments 2B and 2C, our Rhombus and Hexagon conditions). We also tested search for a bottom-lit cube among top-lit ones (their experiment 3A, our Top Black condition) and vice versa (their experiment 3B, our Top White condition), as this pair led to a search asymmetry (Enns & Rensink, 1990a). The Top Black condition is of course quite similar to the Cube condition, but we ran it nonetheless so as to replicate their conditions.

Stimuli were presented on an LCD screen, with subjects seated 75 cm away in a dark room. We ran our experiments in MATLAB, using the Psychophysics Toolbox (Brainard, 1997). The stimuli consisted of a number of display items (the set size), either all distractors (target-absent) or one target and the rest distractors (target-present). Target-present and target-absent displays occurred with equal probability.

Set size was one, six, 12, or 18. Stimuli were randomly placed within selected locations on four concentric circles, plus added positional jitter. The radii of the circles were 4°, 5.5°, 7°, and 8.5° of visual angle (v.a.). Each item had a height of 1° v.a. The cubes were also 1° v.a. in width, while the rhombus and hexagon patterns were somewhat wider (1.5° and 1.2° v.a., respectively). Sample target-present stimuli for the Cube, Rhombus, and Hexagon conditions are shown in Figure 4.

At the beginning of each trial, a fixation cross appeared in the center of the display for 350 ms. The search display then appeared until subjects responded. Subjects indicated with a key press whether or not the display contained a target, and were given auditory feedback. We measured reaction times (RTs) for correct trials. When a subject was incorrect on a given trial, that trial was later repeated. Each subject

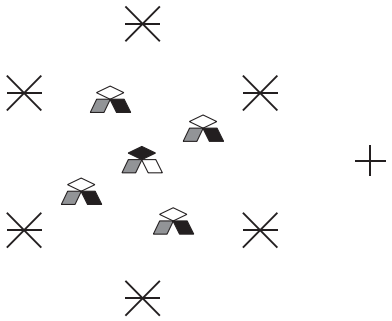

Figure 5. Sample stimulus for the peripheral discriminability task. Numerosity = 5.

completed 40 correct trials (20 target-present and 20 target-absent) for each set size in a given condition. The order of the search conditions was randomized across subjects. For each condition, we computed slopes for RT versus set size. Here we focus on slopes for correct target-present trials.

Peripheral discrimination task

For each of the five search conditions, we ran a corresponding peripheral discrimination task.
Subjects: A different group, of 11 subjects, participated. All had normal or corrected-to-normal vision and received monetary compensation.
Stimuli and procedure: Subjects sat 60 cm away from the display, in a dark room. (The viewing distance was different from that for the search and mongrel tasks due to constraints of the eye-tracker setup. This difference should be inconsequential; we adjusted display items to subtend the same number of degrees of visual angle.)

On each trial, a patch appeared on the display centered at an eccentricity of 7° v.a. Each patch contained a number of distractors, positioned around either a target or a distractor. The subject's task was to respond with a keystroke whether a target was present in the center of the peripheral patch or not. The subjects were instructed that if the target was present, it was always located in the center of the patch; target location was known, so that this task required peripheral discrimination and not visual search.

The surrounding distractors appeared evenly spaced on a notional circle of radius 1.9° v.a. from the target. Thus, the distractors lay within the critical spacing of crowding (Bouma, 1970) and within the range of distances between neighboring items in the search displays. In addition, each peripheral patch was further flanked by a hexagonal array of asterisks (*), evenly spaced on a notional circle of radius 3.8° v.a. from the target. The size of the asterisk was approximately 1° v.a. The circle of asterisks helped to further localize the center of the patch, i.e., the location of the possible target. Each patch contained three, four, or five items (numerosity). That is, on target-present trials, a target was flanked by two, three, or four distractors (see Figure 5).

At the start of each trial, a fixation cross appeared for 350 ms, followed by the stimulus. We took several steps to ensure that the subjects maintained fixation on the central cross throughout each trial. The observers had no prior knowledge about which side (left or right) the patch would appear on, as this varied randomly from trial to trial. We used brief presentation times (180 ms), as is common in peripheral vision tasks (e.g., Balas et al., 2009; Carrasco et al., 1995; Carrasco & Frieder, 1997; Carrasco et al., 1998). This brief time mimics a natural search task, in which a subject would plan and execute a saccade within about 200 ms. Finally, the display was gaze contingent, disappearing if the subject broke fixation. Eye movements were recorded at 240 Hz using a 240-Hz ISCAN RK-464 video-based eye tracker.

Each subject completed 80 trials for each numerosity value in each condition. Half of the trials were target-present. Trials were blocked by condition, with the order of the five conditions randomized for each subject.

In a later experiment, with five new subjects, we also reran the peripheral discrimination experiment with the five classic search conditions from our previous work (Rosenholtz, Huang, Raj et al., 2012). This enabled a more appropriate comparison of the conditions in this article with those classic conditions; the previous study displayed the stimuli at a larger eccentricity (12°) and without the gaze-contingent display. In both experimental setups, the flankers lie within the critical spacing of crowding, and a comparison with the previous study shows that the change in methodology made little difference to the overall results.

Mongrel discrimination task

Subjects: The mongrel discrimination task was carried out by 10 new subjects. All reported normal or corrected-to-normal vision and were paid for their participation.
Stimuli and procedure: For each of the five search conditions, we randomly sampled 10 target-present and 10 target-absent patches from our visual search displays. The patches were 6.4° × 6.4° v.a., centered at an eccentricity of 7° v.a., i.e., the third concentric circle in the search tasks (Figure 6). Although studies have found spatial pooling regions to be more elliptical in shape, elongated radially (Toet & Levi, 1992), these square patches still closely capture the contents of typical pooling regions and are more convenient computationally. We constrained the numerosity of patches to three, four, or five items, to match patch numerosity in the peripheral discrimination experiment.

For each patch, we synthesized 10 new image patches to have approximately the same model image

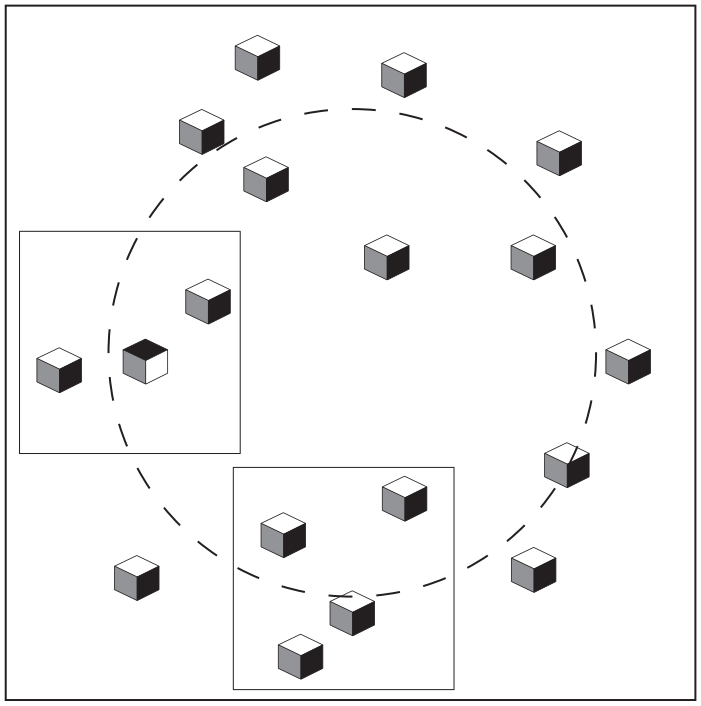

Figure 6. Schematic illustration of sampling of patches from actual search displays, for the mongrel discriminability task.

statistics as the original patch, using Portilla and Simoncelli's (2000) texture synthesis algorithm with four scales, four orientations, and a neighborhood size of nine. This choice of parameter settings leads to measurement of approximately 1,000 image statistics per patch. The algorithm first measures the set of image statistics described earlier in the Texture Tiling Model section. The synthesis algorithm starts with a randomly chosen sample of white noise. It then iteratively adjusts this seed image until it has approximately the same statistics as the original image patch.

The full version of the Texture Tiling Model (Rosenholtz, 2011; Rosenholtz, Huang, & Ehinger, 2012; and similar to the model described by Freeman & Simoncelli, 2011) measures statistics over a number of pooling regions, which overlap and tile the entire visual field. One can synthesize full-field mongrels that share approximately these same local statistics with the original stimulus. Doing so, however, is quite slow. Depending upon the size of the input image, syntheses can take up to 6 h to converge. Generating 1,000 syntheses, as described above, would take a prohibitive amount of processing time. However, we can learn a lot from running the local version of the model, which extracts statistics from only a single pooling region. To understand the implications of this model for tasks like scene recognition, one needs the full-field version, to capture extended structures and larger scale groups for which measurements from multiple pooling regions provide significant additional constraints. For very texture-like search displays like those found here, additional pooling regions provide minimal additional information, and local statistics suffice for making model predictions.

During each trial, a mongrel was presented at the center of the computer screen until subjects responded. Each mongrel subtended 6.4° × 6.4° v.a. at a viewing distance of 75 cm. Subjects were asked to categorize each mongrel according to whether or not they believed the original patch contained a target. They were shown examples of original patches and told those patches had been "jumbled up." They were told that the jumbling could, among other things, flip parts of the stimulus upside down. They were also informed that the patches wrapped around, so that the left edge met the right edge and the top met the bottom. Subjects had unlimited time to freely view the mongrels. As noted in the section describing the Texture Tiling Model, we allowed observers to freely view the mongrels so as to isolate, as much as possible, the low-level factors of interest, i.e., the relatively low-level peripheral representation. Our goal, as in the past, is to judge whether the model—operationalized as mongrel discriminability—can predict peripheral vision *at a glance*, i.e., in less than 200 ms. In particular, we do not equate unlimited mongrel viewing time with unlimited peripheral viewing time of the original display. Additional viewing time may lead to somewhat increased peripheral discriminability relative to that measured here (e.g., Chung & Mansfield, 2009).

We generated, for each condition and set size, 100 mongrels of target-present patches and 100 of target-absent patches. Of these, 20 of each set size were used for training, to familiarize observers with the task and stimuli. Observers received auditory feedback throughout the experiment. It is worth examining the mongrels for each condition, to gain intuitions for the predictions of the model. Figure 7 shows examples for three of the conditions (the Top Black condition produces mongrels like those in the Cube condition, except reflected about the vertical axis; the Top White condition, Figure 15, will be discussed later). In the Cube condition, the model image statistics are sufficiently informative in some cases to even constrain the patch to contain elements that look like cubes. In other words, it appears that *3-D scene properties can be computed from the proposed image statistics*. However, this computation may or may not occur preattentively, and whether such properties underlie easy cube search remains to be seen. We will revisit these issues in discussing our experimental results.

The rhombus condition is more problematic. The weak grouping between the rhombuses causes the mongrels to become disorganized, and due to the lack

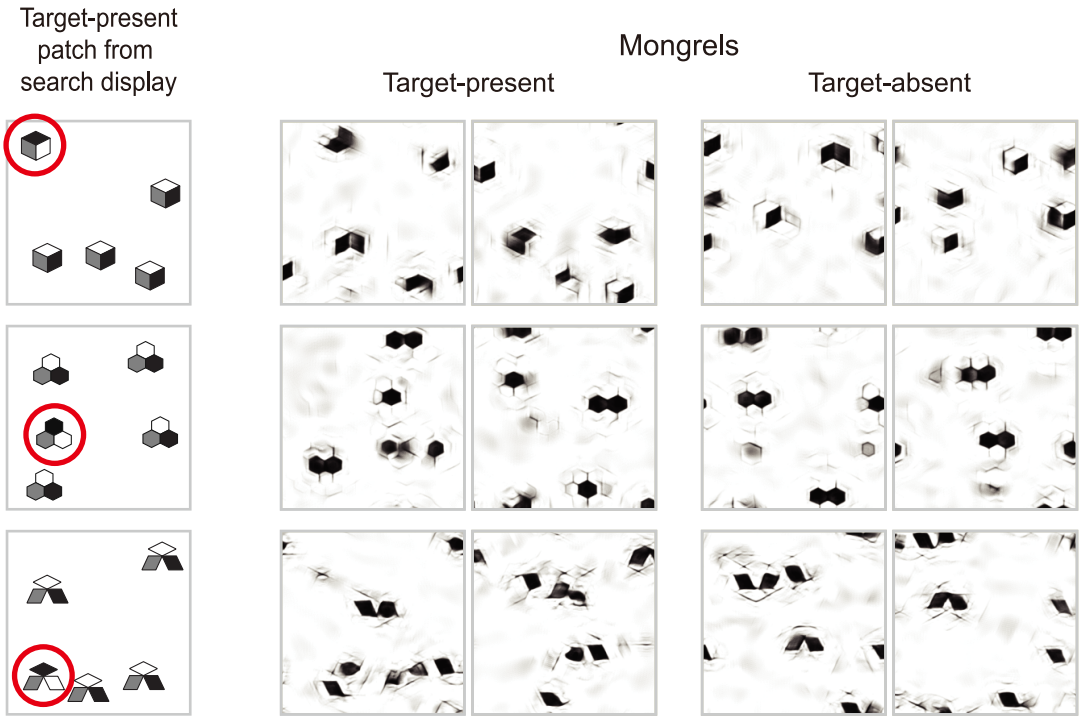

Figure 7. Example mongrels. Left column: Sample target-present patches from the Cube, Hexagon, and Rhombus conditions. Red circle indicates target. Right columns: Two representative mongrels from target-present and target-absent original patches. Observers free-viewed each mongrel for an unlimited time and categorized them according to whether the original patch contained a target.

of organization, the patterns are complex and more poorly represented. This may make the mongrel task in this condition more difficult. The Hexagon condition has almost the opposite problem: Parts of the mongrels appear to contain more grouping structure than is present in the original patches, producing beehive-like structures in which it is difficult to tell if, say, the dark hexagon was on the top or on the side. This confusion may make the Hexagon condition more difficult.

Results

Search difficulty

As is standard in the search literature, we quantify search difficulty as the slope of the best-fit line relating mean RT to the number of items in the display (Figure 8). Only correct target-present trials were included in further analysis (Figure 9). The results are consistent with Enns and Rensink's original study (1990a). Planned comparisons reveal a significant difference between the

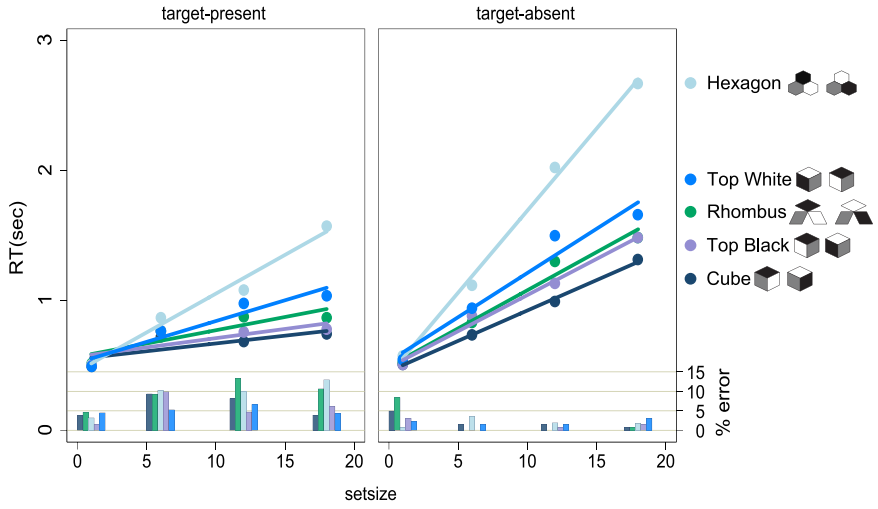

Figure 8. Search results. Reaction time (left axis) and percentage error (right axis, bar plot). Legend shows target on left, distractor on right.

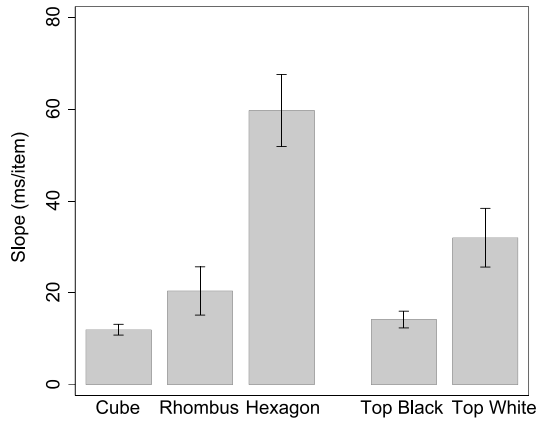

Figure 9. Slopes of RT versus set size in five search experiments. Error bars indicate standard error of the mean.

Cube and Hexagon conditions, $F(1, 20) = 43.2$, $p < 0.01$, and between the Top Black and Top White conditions, $F(1, 20) = 6.02$, $p = 0.023$, but no significant difference between the Cube and Rhombus conditions, $F(1, 20) = 1.35$, $p = 0.26$. While the latter effect was significant in Enns and Rensink's experiments, it is not worth dwelling on this difference. As we shall later see, the interesting analyses come from looking at correlations between a wide range of search conditions and their equivalent peripheral and mongrel tasks.

Peripheral discriminability

Figure 10 shows peripheral discriminability for each condition, where we computed d' separately for each numerosity level. (This figure shows results only for the five Enns & Rensink search conditions. The new peripheral d' data for the classic search conditions of Rosenholtz, Huang, Raj et al., 2012, which we discuss later, appear in Figures 12 and 13). Numerosity has a marginally significant impact on peripheral discriminability (two-way repeated-measures ANOVA), $F(2, 20) = 3.45$, $p = 0.052$, as previously reported for some

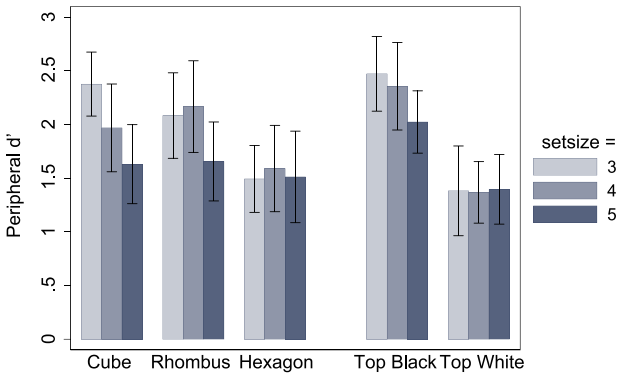

Figure 10. Performance on five peripheral discrimination tasks. Error bars indicate standard error of the mean.

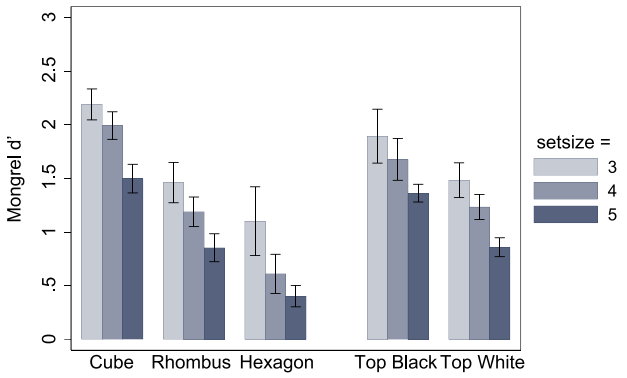

Figure 11. Performance on five mongrel tasks. Error bars indicate standard error of the mean.

crowding conditions (Levi & Carney, 2009; Põder & Wagemans, 2007; Reddy & VanRullen, 2007; Rosenholtz, Huang, Raj et al., 2012; Wertheim et al., 2006). Discriminability decreased with increasing number of items. Moreover, there was a main effect of condition, $F(4, 40) = 5.45$, $p < 0.01$, with post hoc tests revealing that both the Cube and Rhombus conditions were significantly easier then the Hexagon condition and that the Top Black condition was significantly easier than the Top White (Tukey's honestly significant difference [HSD], $p < 0.05$).

Mongrel discriminability

We also measured mongrel discriminability for each condition and each patch numerosity (Figure 11). A repeated-measures ANOVA showed a significant main effect of both set size and condition—set size: $F(2, 18) =$

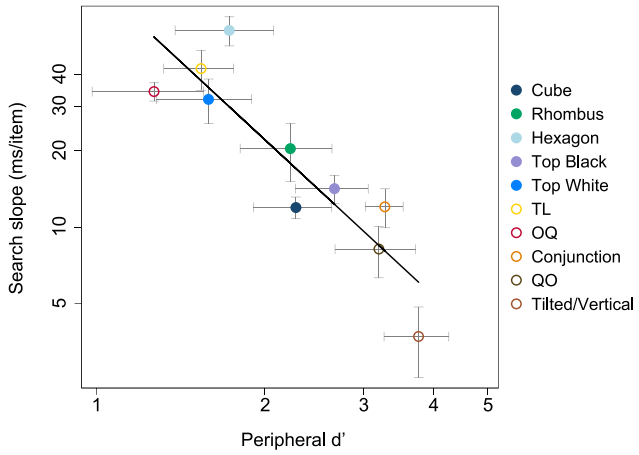

Figure 12. Search efficiency for correct target-present trials versus peripheral discriminability of crowded target-present from target-absent patches. Clearly, there is a strong relationship between visual search difficulty and peripheral discriminability, in agreement with our predictions (error bars indicate standard error of the mean).

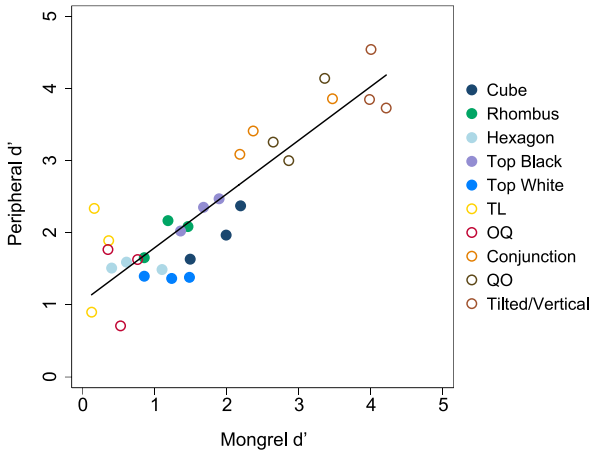

Figure 13. Peripheral d' versus mongrel d'. Different colors represent different conditions, with multiple points corresponding to different patch numerosities.

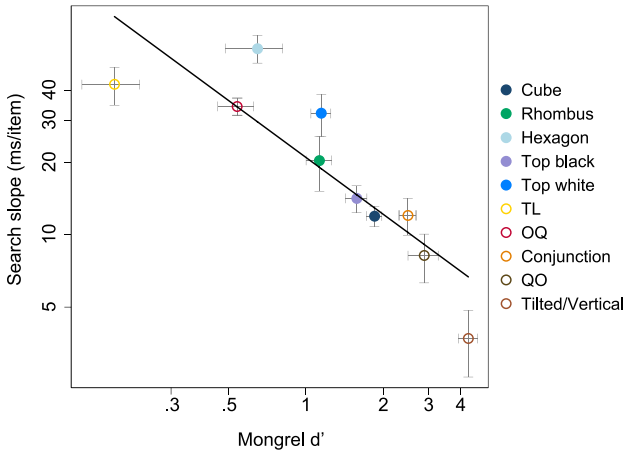

Figure 14. Search efficiency for correct target-present trials plotted against discriminability of mongrels from target-present versus target-absent patches. Again, there is clearly a strong relationship between visual search difficulty and mongrel discriminability, in agreement with our predictions (error bars indicate standard error of the mean).

24.6, $p < 0.01$; condition: $F(4, 36) = 29.3$, $p < 0.01$. Post hoc testing showed a significant decrease in d' for each increase in set size (Tukey's HSD, $p < 0.05$). In addition, cube search was significantly easier than both the Rhombus and Hexagon conditions, and the Top Black condition was significantly easier than the Top White condition (Tukey's HSD, $p < 0.05$). These results are in agreement, in terms of general trend if not always significance, with those of both the search and peripheral discriminability tasks.

Peripheral vision predicts search, and the Texture Tiling Model predicts peripheral vision

Our first, most basic hypothesis was that the information available in peripheral vision is a critical limit on visual search. If this is true, peripheral discriminability of a target-present patch from a target-absent one should predict search performance. Figure 12 plots correct target-present search slope against peripheral d' (across all set sizes), for the five Enns and Rensink conditions and our previous five classic search conditions. There is a clear relationship between these two tasks. We have plotted these results on a log-log plot for convenience; here it appears that the relationship is approximately linear, though we have no a priori reason to expect this. The correlation between log search and log peripheral d' is $R^2 = 0.79$ (significantly different from 0, $p \ll 0.01$). There seems to be nothing particularly special about cube search. Rather, the same function appears to relate peripheral performance (with full attention) to search performance for all 10 conditions.

Having demonstrated a relationship between crowded peripheral identification and search, we next ask how well our model predicts peripheral performance, and thus search. To the extent that our model of peripheral crowding is correct, mongrel discriminability should predict peripheral discriminability. Figure 13 shows the results, where each color corresponds to a given condition, and multiple points of the same color correspond to different patch set sizes. Peripheral discriminability is somewhat easier than mongrel discriminability, particularly for more difficult tasks (intercept = 1.1, greater than 0 with $p < 0.01$). This may mean that the model is missing some information that is available to peripheral vision, or our observers may simply need more training on the mongrel task. Overall, there is a strong relationship between performance on these two tasks ($R^2 = 0.80$, $p < 0.01$).

Finally, we plot search performance versus mongrel d' for all 10 conditions (Figure 14). As with the relationship between peripheral performance and search, we do not at this stage know the expected form of the relationship. However, it is clear that a strong relationship exists ($R^2 = 0.75$, $p \ll 0.01$).

Discussion

These results have several implications. First, many known limits on visual search performance correspond to limits of peripheral vision. This makes sense. Peripheral vision is clearly important for efficiently locating a target. At a minimum, clearly we would have been quite surprised if easy search did not correspond to easy peripheral target discrimination. If the information necessary for efficient search did not survive peripheral vision, that information could not guide

search and we would see very steep search slopes. More interestingly, we also find that difficult search tasks correspond to difficult peripheral discriminations. This need not have been the case. One could imagine finding that fully attended target discrimination is easy in the periphery for all of the tasks, even those corresponding to difficult search tasks. If so, we would clearly need some other mechanism to account for difficult search, such as the need for selective attention to perform those discriminations. But this is not the case; peripheral performance correlates well with search performance overall.

Peripheral vision, in turn, seems well predicted by our relatively low-level model, in which peripheral vision represents its inputs in terms of a particular, rich set of local image statistics. Furthermore, that model itself predicts search efficiency. Again, these are not trivial results. Suppose we had tested a different model. For instance, if we hypothesized that peripheral vision only involved known losses of acuity, the mongrels for that hypothesis would just be slightly blurry versions of the original patches. What would the correlation plots have looked like? Presumably subjects free-viewing these mongrels would be at ceiling for distinguishing target present versus target absent, across all tasks (recall that Figure 2 *exaggerates* the loss of acuity in peripheral vision). Mongrel d' would be consistently high, whereas peripheral and search tasks would show a range of difficulty as in our actual results. Both Figures 13 and 14 would look like a bunch of points scattered about a vertical line. Correlations would be very low. This is not the situation for our model; the hypothesized representation both keeps enough information to support efficient search and throws out enough information to predict difficult peripheral performance and difficult search.

Are the peripheral and mongrel discriminabilities sufficient to support search performance? Of particular concern, subjects in the Enns and Rensink (1990a) experiments seem able to perform even the more difficult tasks while supposedly fixating – albeit with miss rates of around 20% for the highest set sizes. One can get a sense for this task with our stimuli by fixating the center of a display in Figure 4. The peripheral d' for the difficult Hexagon and Top White conditions was around 1.4–1.5. At first glance, these values seem insufficient to support search while fixating. These results would appear puzzling from the point of view of any model of visual search. However, two important points are worth noting. First, we measured peripheral discriminability for short presentation times (and the Texture Tiling Model, similarly, models the information available at a glance). Longer viewing times can improve peripheral performance, at least by a modest amount (e.g., Chung & Mansfield, 2009). More importantly, peripheral d' will vary across the display. Critical spacing varies with eccentricity, and this plus the layout of display items will make some patches more crowded than others. For instance, for an observer fixating the center of Figure 1B, Bouma's law says that four of the 12 display items are actually not crowded in the traditional sense; there are no flankers within the critical spacing of those items. These two factors likely explain the apparent disparity between our peripheral task and Enns and Rensink's (1990a) fixated search task.

Finally, we started by asking whether 3-D scene properties are special, in that they are higher level than many other basic features and yet enable efficient search. Our results do not support this interpretation of Enns and Rensink (1990a). Rather, we have demonstrated an important difference between the Cube condition and both the Rhombus and Hexagon conditions, in terms of 2-D image statistics. Our mongrel discriminability experiment demonstrates that these 2-D image statistics are more informative for distinguishing target-present from target-absent patches in the Cube condition than in the Rhombus or Hexagon condition. The "equivalent" 2-D conditions were not equivalent in terms of this particular set of 2-D image statistics. These results mean there may or may not exist preattentive feature detectors for 3-D scene properties, but *easy cube search, per se, does not provide evidence for their existence*, because there exists a 2-D confound; cube search may be relatively easy because peripheral target-present and target-absent patches can easily be discriminated based on differences in 2-D image statistics. That said, we have had a great deal of success using a relatively low-level model of peripheral vision to predict performance at a glance on a wide range of tasks (Rosenholtz, Huang, & Ehinger, 2012). These results collectively support a late-selection account; if selective attention operates immediately after our modeled peripheral vision mechanisms, we would not expect to do as well at predicting task performance. If selection is relatively late, we would expect that the visual system can make use of the information available in a rich set of 2-D image statistics to preattentively compute certain 3-D scene properties.

It is also clear that Enns and Rensink (1990a) obtained results that our model cannot predict. They found that search for a bottom-lit cube among top-lit cubes (Top Black condition) was easier than vice versa (Top White). This asymmetry reversed when the targets and distractors were turned upside down. One can reason from knowledge of the Texture Tiling Model that it cannot explain this reversal. Flipping the stimuli upside down changes the image statistics measured by the model, but it does not change their informativeness for the task. The statistics measured are predominantly second-order correlations of V1-

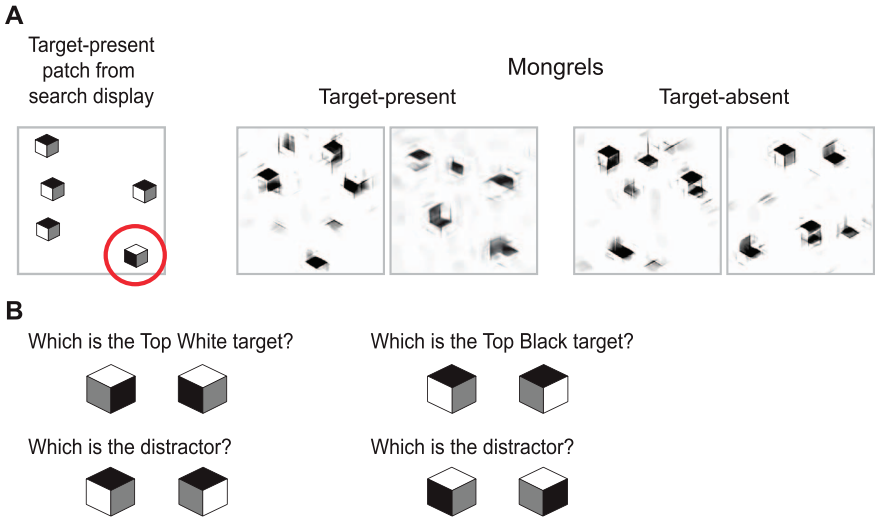

Figure 15. What makes the Top White condition difficult? (A) Two target-present and two target-absent mongrels from the Top White condition, shown with target-present patch (left). The representation in terms of texture statistics seems about as good as for the Cube (or Top Black) condition, shown in Figure 7. (B) It may be difficult to distinguish, e.g., in memory, between a search item and its nearly symmetric reflection about a vertical axis. In the Top White condition, this is a problem. The mongrels predict that the darker faces of the cubes survive peripheral vision well. The subject must keep in mind that the target contains a unique dark face that goes up and to the left. Both target and distractors contain a dark face that goes up and to the right. In the Top Black condition there is no such issue, since the dark top is not easily confusable with the other faces.

like filters computed over each patch. Flipping the stimuli upside down is equivalent to flipping those filters upside down. The upside-down filters are also in the filter set. This means that flipping the stimuli upside down is like relabeling the outputs as coming from filter *i* instead of filter *j*. Mere relabeling in no way changes the informativeness of the statistics. The information available to do the peripheral task is the same, according to the model. Any reversal in asymmetry between the upright and inverted conditions cannot be due to the statistics currently measured in our model.

By similar logic, the asymmetry between the Top Black and Top White conditions is also unlikely to be due entirely to a difference in informativeness of the statistics. These two conditions are nearly (but not quite) a reflection of each other about an oblique line. Such a pair of conditions is likely to have only a small difference in the informativeness of the statistics, and such a small difference is unlikely to have caused the observed search asymmetry. Indeed, the mongrels seem to do about as good a job of representing the Top White stimuli (Figure 15A) as the cube stimuli (Figure 7). Instead, perhaps some higher level factor is responsible for the difficulty in identifying a top white target compared to a top black one. This asymmetry appears in the search, peripheral discriminability, and even mongrel discriminability results, suggesting that we were not completely successful at eliminating higher level effects from our mongrel discriminability experiment. Based on our own subjective experience with these stimuli, it seems that it is difficult to remember, and perhaps even a bit difficult to perceive, the difference between a nearly symmetrical shaded cube and its reflection about a vertical axis. Furthermore, the black faces of the cubes survive well in the mongrels but can be difficult to tell from the gray faces. For the Top White condition this poses a problem, as in the mongrels a good cue to target presence consists of a dark region that shears up and to the left, but the observer must not be fooled by dark regions shearing up and to the right, deriving from both target and distractors. On the other hand, for the Cube and Top Black conditions the near symmetry of the cubes does not pose such a problem. The salient top black diamond in the target is not so confusable with the dark faces in the distractors (Figure 15B). By this account, one can easily imagine parallels to effects of symmetry on search (Wolfe & Friedman-Hill, 1992; Wolfe, Friedman-Hill, Stewart, & O'Connell, 1992), though here the effects appear as well in nonsearch tasks.

Regardless of the cause of the asymmetry in Top Black versus Top White, we have argued in the foregoing that informativeness of the image statistics cannot predict the reversal of the asymmetry when the displays are flipped upside down. What, then, do we make of this reversal? Two possibilities come to mind.

First, the 2-D image statistics in our model may not be not quite right. These statistics were originally hand-selected to capture texture appearance (Portilla & Simoncelli, 2000). Though they have performed well so

far, we would be surprised if they did not require some modification to model peripheral vision. The statistics measured by the visual system perhaps evolved for efficient encoding of task-relevant visual information in the natural world. Perhaps, given a strong prior on lighting from above, it would be efficient for the visual system to preferentially encode certain *2-D patterns*, such as high-luminance regions on top of vertical edges. It is not, however, immediately obvious what change in 2-D statistics would predict the observed asymmetries.

Second, computation of 3-D scene properties may occur preattentively, but such computations may make search *less* efficient, rather than more. After all, our results suggest that performance in the easy Top Black condition is *not* better than expected. Rather, the Top Black condition is *as easy as expected*, based upon the difficulty of a number of 2-D search conditions. Informativeness of the 2-D image statistics predicts both the Top Black and the 2-D search tasks. There is nothing special about easy cube search; what is surprising is the difficulty of the *hard* cube search tasks. This suggests a different interpretation. The visual system must recognize objects in the presence of nuisance variables such as differences in illumination and the resulting differences in shading. In order to gain that invariance, it may encode its input in a way that makes it harder to extract those same nuisance variables. This account puts difficult cube search in the same category as difficult search for the orientation of a shadow (Rensink & Cavanagh, 2004). Drawing or painting a scene requires training precisely because humans have difficulty viewing a 3-D scene and extracting the pixels, even with focal attention.

The involvement of higher level computations in determining search difficulty poses no deep theoretical issues for the Texture Tiling Model. In our model, an early stage represents the visual input using a general-purpose efficient encoding in terms of a fixed, rich set of image statistics. This encoding, while maintaining a great deal of useful information, does entail some loss of information. If the remaining information is insufficient to perform a given task, that task will be difficult. If the information is available to do the task, that task will be easy *unless some later processing stage also throws out or makes inefficient use of information of relevance to the task*. Presumably this sort of information loss happens throughout vision, in service of untangling the perceptual representation to enable simple decision rules for real-world tasks such as recognizing an object regardless of its pose, scale, and lighting (DiCarlo & Cox, 2007). However, this untangling potentially comes at the cost of entangling other, less ecologically relevant tasks, such as distinguishing the amount of light coming from each face of a shaded cube.

Conclusions

Previous work has shown that cube search is easier than similar Hexagon and Rhombus conditions and suggested that the difference is due to the availability of 3-D scene properties in the Cube condition. Our experiments suggest the possibility that the Cube search condition is easier not because differences in 3-D scene properties make search especially easy, but rather because a rich set of 2-D image statistics is more useful for distinguishing between whether a peripheral patch contains a particular target cube than a target hexagon or rhombus pattern. The informativeness of this same set of image statistics predicted classic search results (Rosenholtz, Huang, Raj et al., 2012) as well as the results of the Cube, Hexagon, and Rhombus conditions. Viewed in this way, cube search is not especially easy, compared with other 2-D patterns, but rather is just as easy as expected from the informativeness of the image statistics.

This is not to say that 3-D scene properties are irrelevant to the difficulty of search tasks. Rather, the 3-D interpretation of shaded cubes may actually make search more difficult. While easy cube search is just as easy as expected, difficult search for a cube lit from a common lighting direction is more difficult than one would predict from the informativeness of the image statistics.

An important benefit of our model-driven approach is that it allows us to ask and answer whether cube search is easier or more difficult than expected based on performance on 2-D tasks. Rather than speculating on what might be equivalent 2-D stimuli, we can quantify, according to the model, the informativeness of relatively low-level cues for the different conditions. This has allowed us to make sense of whether cube search is unexpectedly easy or difficult, which was hard through behavioral experiments alone.

Our results provide further evidence for our hypothesis (Balas et al., 2009; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj et al., 2012) that peripheral vision encodes its inputs in terms of a particular, rich set of image statistics. The loss of information inherent in such a representation predicts peripheral discrimination across a range of tasks. Furthermore, inability to peripherally discriminate between target-present and target-absent patches predicts difficulty finding the target, for a range of search conditions. The loss of information in peripheral vision represents a critical way in which vision is not the same everywhere.

Keywords: visual search, image statistics, summary statistics, mongrel, Texture Tiling Model, Feature Integration Theory, familiarity, 3-D shape, lighting direction, peripheral vision

Acknowledgments

This work was funded by NIH-NEI EY021473 to RR and by NSFC 61305096 to XZ. XZ gratefully acknowledges financial support from China Scholarship Council.

Commercial relationships: none.
Corresponding author: Ruth Rosenholtz.
Email: rruth@mit.edu.
Address: Department of Brain and Cognitive Sciences, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.

References

Allport, A. (1993). Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. In E. Meyer & S. Kornblurn (Eds.), *Attention and performance XVI: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 182–218). Cambridge, MA: MIT Press.

Anstis, S. (1998). Picturing peripheral acuity. *Perception, 27,* 817–825.

Balas, B. J. (2006). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research, 46*(3), 299–309.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12):13, 1–18, http://www.journalofvision.org/content/9/12/13, doi:10.1167/9.12.13. [PubMed] [Article]

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226,* 177–178.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Carrasco, M., Evert, D. L., Chang, I., & Katz, S. M. (1995). The eccentricity effect: Target eccentricity affects performance on conjunction searches. *Perception & Psychophysics, 57*(8), 1241–1261.

Carrasco, M., & Frieder, K. S. (1997). Cortical magnification neutralizes the eccentricity effect in visual search. *Vision Research, 37*(1), 63–82.

Carrasco, M., McLean, T. L., Katz, S. M., & Frieder, K. S. (1998). Feature asymmetries in visual search: Effects of display duration, target eccentricity, orientation and spatial frequency. *Vision Research, 38*(3), 347–374.

Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 24*(2), 673–692.

Chung, S. T. L., & Mansfield, J. S. (2009). Contrast polarity differences reduce crowding but do not benefit reading performance in peripheral vision. *Vision Research, 49*(23), 2782–2789.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341.

Di Lollo, V., Kawahara, J. I., Zuvic, S. M., & Visser, T. A. (2001). The preattentive emperor has no clothes: A dynamic redressing. *Journal of Experimental Psychology: General, 130*(3), 479–492.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 9*(2), 111–118.

Enns, J. T., & Rensink, R. A. (1990a, February 9). Influence of scene-based properties on visual search. *Science, 247*(4943), 721–723.

Enns, J. T., & Rensink, R. A. (1990b). Sensitivity to three-dimensional orientation in visual search. *Psychological Science, 1*(5), 323–326.

Enns, J. T., & Rensink, R. A. (1991). Preattentive recovery of three-dimensional orientation from line drawings. *Psychological Review, 98*(3), 335–351.

Erkelens, C. J., & Hooge, I. T. C. (1996). The role of peripheral vision in visual search. *Journal of Videology, 1,* 1–8.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1201.

Geisler, W. S., & Chou, K. L. (1995). Separation of low-level and high-level factors in complex tasks: Visual search. *Psychological Review, 102*(2), 356.

Geisler, W. S., Perry, J. S., & Najemnik, J. (2006). Visual search: The role of peripheral information measured using gaze-contingent displays. *Journal of Vision, 6*(9):1, 858–873, http://www.journalofvision.org/content/6/9/1, doi:10.1167/6.9.1. [PubMed] [Article]

Gheri, C., Morgan, M. J., & Solomon, J. A. (2007). The relationship between search efficiency and crowding. *Perception, 36*(12), 1779–1787.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene

analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Lettvin, J. Y. (1976). On seeing sidelong. *The Sciences, 16*(4), 10–20.

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research, 48*(5), 635–654.

Levi, D. M., & Carney, T. (2009). Crowding in peripheral vision: Why bigger is better. *Current Biology, 19*(23), 1988–1993.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9–16.

Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 411–422). Cambridge, UK: Cambridge University Press.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*(7031), 387–391.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision, 8*(3):4, 1–14, http://www.journalofvision.org/content/8/3/4, doi:10.1167/8.3.4. [PubMed] [Article]

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision Research, 49*(10), 1286–1294.

Ostrovsky, Y., Cavanagh, P., & Sinha, P. (2005). Perceiving illumination inconsistencies in scenes. *Perception, 34,* 1301–1314.

Palmer, J., Ames, C. T., & Lindsey, D. T. (1993). Measuring the effect of attention on simple visual search. *Journal of Experimental Psychology: Human Perception and Performance, 19*(1), 108–130.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*(10), 1227–1268.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(12):12, 1136–1169, http://www.journalofvision.org/content/4/12/12, doi:10.1167/4.12.12. [PubMed] [Article]

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11*(10), 1129–1135.

Põder, E., & Wagemans, J. (2007). Crowding with conjunctions of simple features. *Journal of Vision, 7*(2):23, 1–12, http://www.journalofvision.org/content/7/2/23, doi:10.1167/7.2.23. [PubMed] [Article]

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–70.

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature, 331*(6152), 163–166.

Reddy, L., & VanRullen, R. (2007). Spacing affects some but not all visual searches: Implications for theories of attention and crowding. *Journal of Vision, 7*(2):3, 1–17, http://www.journalofvision.org/content/7/2/3, doi:10.1167/7.2.3. [PubMed] [Article]

Rensink, R. A., & Cavanagh, P. (2004). The influence of cast shadows on visual search. *Perception, 33*(11), 1339–1358.

Rohaly, A. M., Ahumada, A. J., Jr., & Watson, A. B. (1997). Object detection in natural backgrounds predicted by discrimination performance and models. *Vision Research, 37*(23), 3225–3235.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research, 39*(19), 3157–3163.

Rosenholtz, R. (2011). What your visual system sees where you are not looking. In B. E. R. Pappas and T. N. Pappas (Eds.), *SPIE: Human Vision and Electronic Imaging, XVI,* 7865, 786510, doi:10.1117/12.876659.

Rosenholtz, R. (2014). Texture perception. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. NN–NN). Oxford, UK: Oxford University Press, doi:10.1093/oxfordhb/9780199686858.013.058.

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,* 1–15.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision, 12*(4):14, 1–17, http://www.journalofvision.org/content/12/4/14, doi:10.1167/12.4.14. [PubMed] [Article]

Sun, J., & Perona, P. (1996). Early computation of shape and reflectance in the visual system. *Nature, 379*(6561), 165–168.

Toet, A., & Levi, D. M. (1992). The two-dimensional

shape of spatial interaction zones in the parafovea. *Vision Research, 32*(7), 1349–1357.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*(4–8), 411–443.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*(1), 107–141.

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*(1), 507–545.

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience, 16*(1), 4–14.

Verghese, P., & Nakayama, K. (1994). Stimulus discriminability in visual search. *Vision Research, 34*(18), 2453–2467.

Vlaskamp, B. N., Over, E. A., & Hooge, I. T. C. (2005). Saccadic search performance: The effect of element spacing. *Experimental Brain Research, 167*(2), 246–259.

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception & Psychophysics, 56*(5), 495–500.

Watson, A. B. (1993). Visual optimization of DCT quantization matrices for individual images. *Proceedings of American Institute of Aeronautics and Astronautics Computing in Aerospace, 9,* 286–291.

Wertheim, A. H., Hooge, I. T. C., Krikke, K., & Johnson, A. (2006). How important is lateral masking in visual search?. *Experimental Brain Research, 170*(3), 387–402.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*(2), 202–238.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419–433.

Wolfe, J. M., & Friedman-Hill, S. R. (1992). On the role of symmetry in visual search. *Psychological Science, 3*(3), 194–198.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., & O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology, 18*(1), 34–49.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*(6), 495–501.

Zetzsche, C., Barth, E., & Wegmann, B. (1993). The importance of intrinsically two-dimensional image features in biological vision and picture coding. In A. B. Watson (Ed.), *Digital images and human vision* (pp. 109–138). Cambridge, MA: MIT Press.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article]