Proc. SPIE: Human Vision and Electronic Imaging, XVI, B.E. Rogowitz & T. N. Pappas, Eds., San Francisco, CA, 2011.

What your visual system sees where you are not looking

Ruth Rosenholtz*

Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

What is the representation in early vision? Considerable research has demonstrated that the representation is not equally faithful throughout the visual field; representation appears to be coarser in peripheral and unattended vision, perhaps as a strategy for dealing with an information bottleneck in visual processing. In the last few years, a convergence of evidence has suggested that in peripheral and unattended regions, the information available consists of local summary statistics. Given a rich set of these statistics, many attributes of a pattern may be perceived, yet precise location and configuration information is lost in favor of the statistical summary. This representation impacts a wide range of visual tasks, including peripheral identification, visual search, and visual cognition of complex displays. This paper discusses the implications for understanding visual perception, as well as for imaging applications such as information visualization.

Keywords: Summary statistics, Texture Tiling model, peripheral vision, visual search, crowding, visualization

1. INTRODUCTION

1.1 A bottleneck in vision

Vision is an active process: we repeatedly move our eyes to seek out objects of interest and explore our environment. Nonetheless, a fundamental constraint on our performance of visual tasks is what we can see in a single glance. If an alert "pops out" and draws our attention, we can easily and quickly notice it even if we are not looking right at it. If a driver can quickly glance at her GPS system and tell that she is approaching a left turn, she will more effectively use her GPS than if comprehending the display requires several glances. A complex diagram, like a subway map, is unlikely to be fully comprehended at a glance, but in a well designed map the viewer has adequate information for planning his next glance, and for piecing together his route.

The question of what our visual systems can perceive in a glance would be boring, except that processing is not uniform throughout the visual field. Some regions of the visual field, most notably the fovea, are rendered more faithfully than others. As a result, the information available in a particular glance typically differs from the information available in the next. This phenomenon is precisely what forces us to glance around to begin with. Furthermore, we are far from optimal at piecing together information from multiple glances into a coherent whole,[1,2,3] despite the fact that we *feel* like we have a unified, stable percept of our visual world.

One piece of evidence for visual processing being spatially non-uniform comes from visual search. We often search inefficiently for a target item among other, distractor items, even when the target is quite easily distinguishable from any individual distractor.[4,5] Figure 1a shows an example: search for a light square among light triangles and dark squares. Vision must not be the same everywhere across the visual field; if it were, the easy discriminability of target from distractors should predict easy search.

Additional evidence comes from the phenomenon of change blindness.[2,3] In a glance, we can get the gist of an image, such as a complex natural scene.[6] We feel as if we are aware of a rich representation of the image. However, when probed it becomes clear that the details are murky. If the two images in Fig. 1b are shown as successive frames in a movie, it is easy to see the difference between the two. But if we remove motion as a cue, here by putting the two images side by side, it becomes difficult to spot the difference. Again, if vision were equally faithful everywhere, it should be easy to detect this change, as once we notice it the change is clearly visible.

Numerous visual illusions also have a component due to the non-uniformity of visual processing. For some illusions, the illusory effect exists predominantly in the periphery; e.g. in the Pinna-Gregory illusion[7] (Fig. 2a), the concentric circles seem to intersect in the periphery. For other illusions, such the bistable Necker cube and Schroeder stairs (Fig. 2b), the percept depends upon where one points ones' eyes or attention.[1,8,9] The non-uniformity of vision is likely also

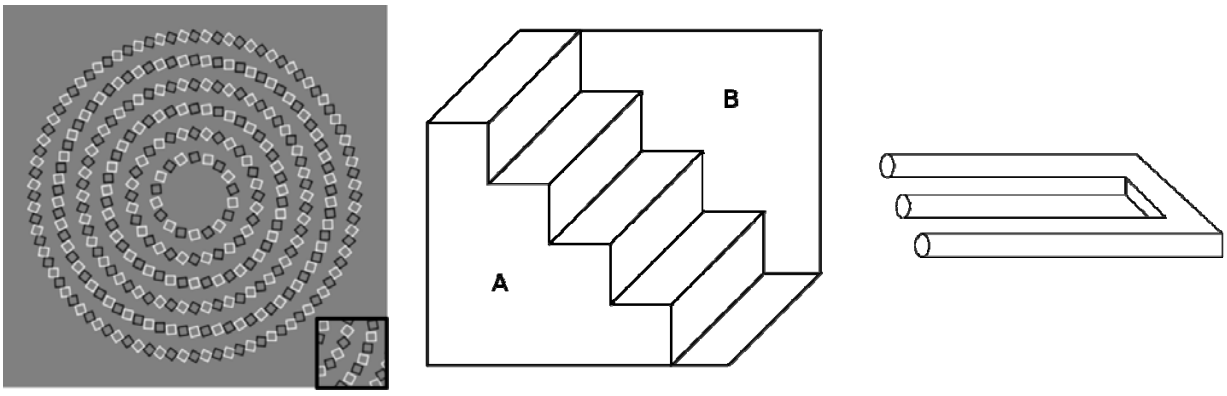
* E-mail: rruth@mit.edu, URL: http://persci.mit.edu/people/rosenholtz, Telephone: 1 617 324-0269

(a) (b)

Figure 1. Evidence for an information bottleneck in vision. (a) Search for a light square among dark squares and light triangles is relatively inefficient. However, when we look at the light square, it is clearly discriminable from both types of distractors. This puzzling behavior implies that vision is not the same throughout the visual field. The foveal discriminability of an individual target from individual distractors is poorly predictive of search difficulty because peripheral vision is not like focal vision. (b) Though we feel like we have at all times a rich representation of the visual world, explicitly probing this knowledge, as with the change-detection task shown here, demonstrates that the details are murky where we are not looking. (If reading this paper electronically, it is recommended that one view Fig. 1b at increased zoom.)

responsible for our difficulty determining the impossibility of a figure such as a blivet, a.k.a. a devil's fork (Fig. 2c); it is difficult to simultaneously perceive that the left side of the fork has 3 tines, while the right side of the fork has only two.[1] The percept of such impossible figures also depends upon where one points one's eye and/or attention.[10]

Visual search and change blindness,[11] in particular, as well as degraded performance at dual tasks,[12] have been taken as evidence of an information bottleneck in vision. The idea is that to accommodate this bottleneck, information is more coarsely encoded in parts of the visual field where we are "not looking." "Not looking" could mean not pointing our eyes at a given region (i.e. not foveating), not attending to that region, or diffusely attending across a broader region. For this paper, we focus on "not foveating." In the Discussion (Sec. 4), we briefly examine whether a similar strategy might account for reported differences between attended and unattended vision.



(a) (b) (c)

Figure 2. What you see depends upon where you look. (a) Pinna-Gregory illusion[7]. The circles are actually concentric, but appear to intersect in interesting ways. The illusion is nearly gone near the fovea (see small patch, inset); (b) Schroeder stairs. Fixation location can bias perception of whether "A" or "B" is in front. (c) Looking to the right, this blivet appears to have two tines, and the left side is ambiguous. This may be why it is difficult to tell that the figure is impossible.

Figure 3. Visual crowding. The "A" on the left is easy to recognize, if it is large enough, whereas the A amidst the word "BOARD" can be quite difficult to identify. This cannot be explained by a mere loss of acuity in peripheral vision.

Peripheral vision is, as a rule, worse than foveal vision, and often much worse. Only a finite number of nerve fibers can emerge from the eye, and rather than providing uniformly mediocre vision, the eye trades off sparse sampling in the periphery for sharp, high resolution foveal vision. If we need finer detail (for example for reading), we move our eyes to bring the fovea to the desired location. This economical design continues into the cortex: the cortical magnification factor expresses the way in which cortical resources are concentrated in central vision at the expense of the periphery.

However, acuity loss is not the entire story, as made clear by the visual phenomena of crowding. An example is given in Fig. 3. A reader fixating the central "+" will likely have no difficulty identifying the isolated letter on the left. However, that same letter can be difficult to recognize when it is flanked by additional letters, as shown on the right. This effect cannot be explained by a simple loss of acuity, as the reduction in acuity necessary to cause flankers to interfere with the central target on the right would also completely degrade the isolated letter on the left.

It is crucial, in order to understand vision, to characterize the information available where we are not looking. Even by a conservative estimate, where we are not looking takes up 99% of the visual field. Furthermore, the representation outside focal attention is crucial to many visual tasks: it guides eye movements, enables quick judgments about, e.g., the gist of a scene,[6] and determines what tasks we can do without the difficult task of piecing together information across a series of fixations. Section 1.2 gives intuitions behind our proposed representation[13] in peripheral vision. Section 2 reviews evidence that such a representation underlies visual crowding as well as visual search performance.

1.2 A strategy for getting through the bottleneck

Given that peripheral vision involves a loss of information, what information should be retained? Imagine representing a patch in the periphery by a finite set of numbers. These numbers could be the firing rates of a finite set of neurons or some other low-dimensional representation. More concretely, suppose that we wanted to represent the image in Fig. 4a with just 1000 numbers. We could coarsely subsample this patch down to a 32x32 array of pixel values, using standard filtering and sampling techniques. This is akin to peripheral subsampling in the retina, and leads to a representation like Fig. 4b. Another option would be to convert Fig. 4a to a wavelet-like representation like that in early visual cortex (V1) – local orientation at multiple scales– and then select the most useful 1000 coefficients. Essentially, if each coefficient corresponds to a potential "neuron", then one can think of choosing the 1000 neurons with the highest

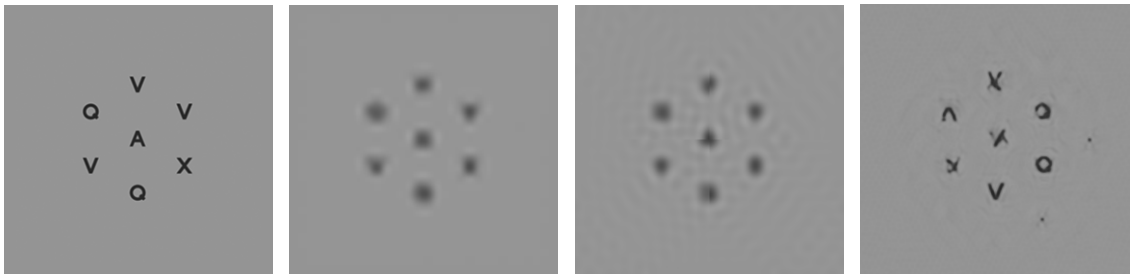


Figure 4. A demo to provide insight into possible coarse encoding strategies for peripheral vision. (a) An original image, to be viewed peripherally. Suppose, hypothetically, that we want to represent this image with only 1000 numbers. (b) Subsampling to reduce to a 32x32 image. Clearly this would be a poor representation. One can tell that the original stimulus consisted of 7 items in an array, but we have no idea that those items were made up of lines, nor that they formed letters. (c) Representation by local orientation at multiple scales, as in early visual cortex (V1), followed by reduction to 1000 numbers leads to a similarly poor result. This encoding used the discrete cosine transform; using more biologically plausible wavelets leads to similar results. (d) For the same 1000 numbers, one can encode a whole bunch of summary statistics, e.g.: the correlation of responses of V1-like cells across location, orientation, and scale; phase correlation; marginal statistics of the luminance; and autocorrelation of the luminance. Here we visualize the information available from those statistics by synthesizing a new "sample" with the same statistics as those measured from (a), using a technique (and statistics) from Portilla & Simoncelli[14]. This encoding captures much more useful information about the original stimulus.

expected firing rates. This leads to a representation like that in Fig. 4c. Both of these strategies discard the high spatial frequencies, which makes it impossible to tell much about the resulting blobs other than their locations.

Suppose, however, that it is valuable to know that the objects are letter-like. Is there a way to encode this visual quality while staying within our hypothetical 1000 number limit? We might instead measure a rich set of summary statistics. In particular: the marginal distribution of luminance; luminance autocorrelation; correlations of the magnitude of responses of oriented V1-like wavelets across differences in orientation, neighboring positions, and scale; and phase correlation across scale. These are summary statistics that have been shown to do a good job of capturing texture appearance.[14,15] Such a representation can capture detailed information about the appearance of the objects at the expense of increased positional uncertainty. Figure 4d shows a sample of texture synthesized[14] to have approximately the same high-order image statistics as found in Fig. 4a. The results are intriguing. Patches synthesized in this way contain evenly spaced arrays of letter-like objects. The exact details and locations are somewhat jumbled, but the model captures the "look" of the original in important ways. These properties are reminiscent of some of those found in peripheral vision and exemplified by crowding. Indeed, recent research on crowding has suggested that the representation in peripheral vision consists of summary statistics computed over local pooling regions.[13,16,17,18]

Does it make sense for peripheral vision to retain statistical information about a pattern's appearance, while losing the arrangement of the pattern elements? The answer may come from considering the different roles played by foveal and peripheral vision. Foveal vision contains powerful machinery for object recognition, but covers a tiny fraction of the visual field. A major role of peripheral vision, by comparison, is to monitor a much wider area, looking for regions that appear interesting or informative, in order to plan eye movements. Take as an example the task of visual search. The task is to look for a target, say, the letter O. At each instant, the subject must quickly survey the entire visual field, seeking out regions worthy of further examination. If the informational bottleneck has reduced everything to fuzzy blobs (Figs. 4bc), then there is no way to choose among the blobs. However, if one at least knows that a particular patch contains O-like stuff – information which is available in Fig. 4d – then an eye movement can be launched in the right direction, and the search process can proceed.

Furthermore, a number of visual tasks inherently require statistical information, for which such a representation might be useful. "Preattentive" texture segmentation involves the rapid detection of a boundary between two texture regions. This process has long been thought of, in both human and computer vision, as involving statistical inference,[19,20,21,22] as has texture classification.[23,24] The phenomenon of "popout," in which an unusual item seems to draw our attention and thus be easy to search for, has been characterized as outlier detection.[25] Recent work has suggested that skew of both the luminance histogram and sub-band filter outputs serve as a cue for perception of shininess of a material.[26] Finally, in deciding where to forage for berries, the visual system might make use of statistical properties such as the mean size of the berries, and in fact humans can estimate such properties.[27,28,29]

We note two distinctions between our proposed representation and that popular in set perception.[27,28,29] The summary statistics we refer to are statistics of the *stuff* in each local patch of the visual input. The set statistics are more often about *things*, e.g. the mean size of a number of elements. Furthermore, our argument, above, that local summary statistics might actually provide a useful means of getting around a bottleneck in vision, hinges on the use of a very rich set of summary statistics. As shown in Fig. 4d, this rich set of summary statistics – *far* more information than merely a few ensemble statistics like mean size and mean orientation – captures much of the appearance of the original patch.

2. A SUMMARY STATISTIC REPRESENTATION PREDICTS VISUAL CROWDING AND PERFORMANCE AT VISUAL SEARCH

2.1 A testable hypothesis for representation in early vision

2.1.1 What summary statistics?

As suggested above, we hypothesize that within each local pooling region, the visual system represents its input by a rich set of summary statistics. Though further investigation will be required to pinpoint exactly what statistics are involved, previous work has suggested that a good initial guess is the statistics used to generate Fig. 4d. These summary statistics were previously suggested for capturing texture appearance for purposes of texture synthesis,[14] and include: marginal statistics of luminance and color; autocorrelation; correlations of responses of V1-like cells across location, orientation, and scale; and phase correlation across scales. See Ref. 14 for more details.

Why are these summary statistics a good initial choice? Certainly they seem quite plausible as a visual system representation. Early stages of standard feed-forward models of object recognition typically measure responses of

oriented, V1-like feature detectors, as does our model. They then build up progressively more complex features by looking for co-occurrence of simple structures over a small pooling region.[30,31] These co-occurrences, computed over a larger pooling region, can approximate the correlations computed by our model.

Second, they appear to be quite close to sufficient. Balas[15] showed that observers are barely above chance at parafoveal discrimination between a grayscale texture synthesized with this set of statistics and an original patch of texture. More recent results have shown a similar sufficiency of these summary statistics for capturing the appearance of real scenes. Researchers synthesized full-field versions of natural scenes. These syntheses were generated to satisfy constraints based on local summary statistics in regions that tile the visual field and grow linearly with eccentricity (see Sec. 2.1.2). When viewed at the appropriate fixation point, observers had great difficulty discriminating real from synthetic scenes.[32] Though both of these results indicate only sufficiency of the proposed statistics, that is impressive nonetheless; much information has been thrown away, and yet observers have difficulty telling the difference.

Finally, significant subsets of the proposed summary statistics are also *necessary*. If a subset of statistics is necessary, then textures synthesized without that set should be easily distinguishable from the original texture. Balas[15] has shown that observers become much better at parafoveal discrimination between real and synthesized textures when the syntheses do not make use of either the marginal statistics of luminance, or of the correlations of magnitude responses of V1-like oriented filters.

There is less work we can draw on to say how color should be represented. This is not an issue for the crowding and search work described in Secs. 2.2 and 2.3, as those stimuli are grayscale. It seems likely that the visual system computes summary statistics in several color channels, and perhaps also computes some sort of correlations between those channels. More research is required to figure out how color is represented. For the purpose of the demos in Secs. 3 and 4, we first used independent components analysis[33] to split the image into three color bands. We measured statistics in each of these bands independently, as in the grayscale case. Within each local pooling region we also measured the covariance between the three color bands.

2.1.2 What pooling regions?

What do we know about the pooling regions over which the summary statistics are computed? Work in visual crowding suggests that they grow linearly with eccentricity – i.e. with distance to the center of fixation -- with a radius of approximately 0.4 to 0.5 the eccentricity. This has been dubbed "Bouma's law," and it seems to be invariant to what is actually in the stimulus.[18] The pooling regions also tend to be elongated radially outward from fixation. Note that there is no discontinuity in this representation; in principle, even though we set out to model representation where one is "not looking," the representation we describe could be a continuous representation throughout the visual field. One possible caveat is that that pooling region is unlikely to be of size 0 at fixation, which implies some deviation from Bouma's law in the fovea. Presumably overlapping pooling regions tile the entire visual input. We call our model of visual representation in terms of the hypothesized "texture" statistics, computed over local pooling regions that tile the visual input in this fashion, the *Texture Tiling* model.

The pooling regions may be fixed in retinal coordinates, or it may be possible for them to shift to a limited degree. The visual system is almost certainly limited in the overall number or density of pooling regions; if it were not, there would be no compression, no loss of information in this scheme. For the purposes of studying visual crowding and visual search (Secs. 2.2 and 2.3), we have designed our experiments so that we can examine the information available in a single pooling region, and show that it predicts task performance. Doing so allowed us to study these phenomena in advance of clear answers on questions of pooling region layout. For the demos in Secs. 3 and 4, we used a pooling region radius of 0.4 × eccentricity. For overlap, we somewhat arbitrarily assumed that neighboring pooling regions at the same eccentricity overlapped over approximately 46% of their area. Radially, neighboring pooling regions overlapped such that the larger, more eccentric regions overlapped approximately 58% of the area of their less eccentric neighbors, whereas the less eccentric regions covered approximately 26% of the area of their more eccentric neighbors.

With assumptions in this ballpark, the number of summary statistics measured is typically only modestly less than the number of pixels, *N*, in the original image. This sounds like a poor compression ratio, but – the demo in Fig. 3 aside – it is not the right comparison. As both the fields of human and computer vision know, little inference can be done with pixels, and the visual system no doubt does not have the option of passing anything like pixels through the bottleneck. For inference, one wants to measure local orientation at multiple scales, as in V1, and then piece them together into more complex and useful structures like complex cells, co-occurrences of horizontal and vertical, and so on. The correct comparison, then, is between our hypothesized representation and a full pyramidal representation of outputs of feature

detectors at 4 orientations, 2 phases, and 4 scales, plus co-occurrences between pairs of those filter outputs. Depending upon how many pairs of co-occurrences are computed, the number of measurements in the uncompressed scheme can range from about 10*N* (if only simple cell responses get through the bottleneck) through at least 90*N* (same pairwise co-occurrences as in our model, but computed at every location rather than pooled over each region). This suggests that the hypothesized representation does in fact achieve a reasonable degree of compression over more obvious alternatives.

Section 2.2 reviews evidence that the Texture Tiling model predicts results from visual crowding. Section 2.3 revisits visual search with this model in hand.

2.2 Visual crowding

In looking for evidence of a visual representation in terms of summary statistics, one should look where vision seems to be broken as a result of the loss of information. Visual crowding, described above, provides an obvious choice, as it demonstrates significant loss of information in peripheral vision.

In order to test whether Texture Tiling predicts visual crowding, we first ran a bunch of crowding tasks in which observers had to indicate which of 4 target letters was present in the middle of an array. Flankers were either other subsets of letters, curves or bars, squiggly lines, or pictures of other objects like a toaster or bike lock. The array was presented at 14 deg eccentricity. The stimuli were designed so that the target and all flankers fell within a single Bouma's law pooling region. We aimed for tasks with a range of difficulty, to give our model something to predict.

How, does one test the model? Suppose for a particular crowding condition the target letters were from the set {F, E, L, T}. One could measure the summary statistics for all the experimental stimuli, i.e. for each peripheral array. Then one could ask how discriminable are the "F"-target stimuli from the "E"-target stimuli, and so on, based on these summary statistics. If our model of peripheral vision is correct, then the discriminability based on the summary statistics should predict performance on crowding tasks, over a wide range of tasks.

One could, of course, design a computer vision algorithm to measure the statistical discriminability. However, this task is effectively a pattern discriminability task, and the best pattern recognizers are humans. Can we not get humans to do the pattern discriminability for us, rather than relying on computer vision? We do this by making use of texture synthesis[14] to generate new images that share approximately the same summary statistics as each original stimulus. We call these images "mongrels." If we can sample the space of images sharing a given set of summary statistics, we can

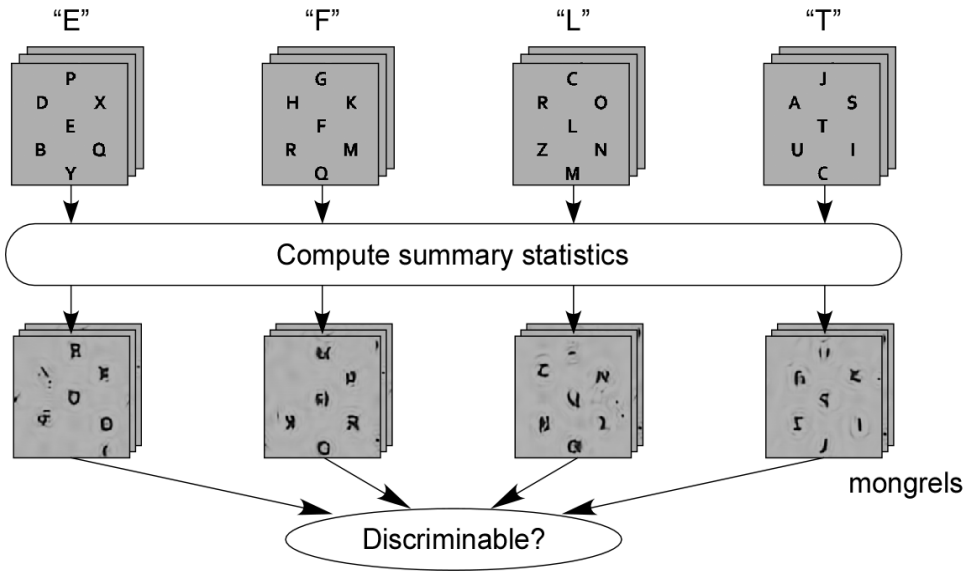


Figure 5. For each condition, stimuli fall into four classes, based upon their central target (top row). For each stimulus (row 2) we measure summary statistics, and then generate "mongrels" – textures with approximately the same summary statistics (row 3). Subjects can then view these mongrels, and classify them into 4 categories corresponding to the 4 target types. This methodology allows us to put a human "in the loop" to better measure the discriminability of these 4 classes based upon our hypothesized representation. Essentially, it gives us a measure of the inherent difficulty in doing a crowding task if the only information available is the measured summary statistics. As Fig. 6 shows, this inherent difficulty is predictive of crowding performance.

effectively visualize the information available in those statistics: the ambiguities and confusions inherent in the representation. Subjects view these mongrels in the fovea, and for unlimited time. We want, as near as possible, for the only information loss to be in going from the original image to the summary statistic representation, so that we can study what task performance is possible with that representation. The subject's task is to discriminate between the 4 target classes (Fig. 5; see Ref. 13 for more details). Again, if our model of peripheral vision is correct, then the discriminability of the mongrels should predict performance on crowding tasks, over a wide range of tasks.

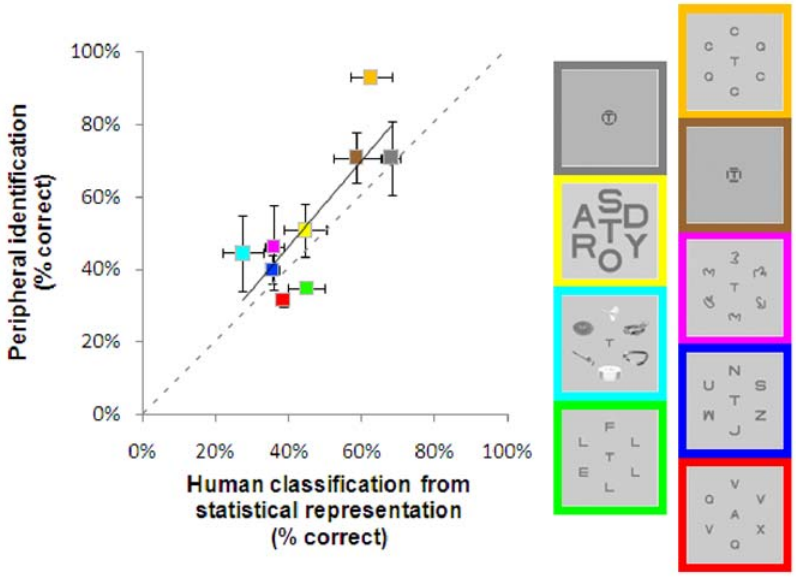


Figure 6. Results of crowding experiments, from Balas et al[13]. Each square represents a different condition, as shown color-coded on the right. The y-axis indicates performance identifying the central target letter in the crowded array. Chance is 25%. The x-axis shows performance discriminating between the four possible targets based upon the summary statistic representation, i.e. based upon synthesized "mongrels" of each stimulus patch. This statistical discriminability is quite predictive of crowded letter identification.

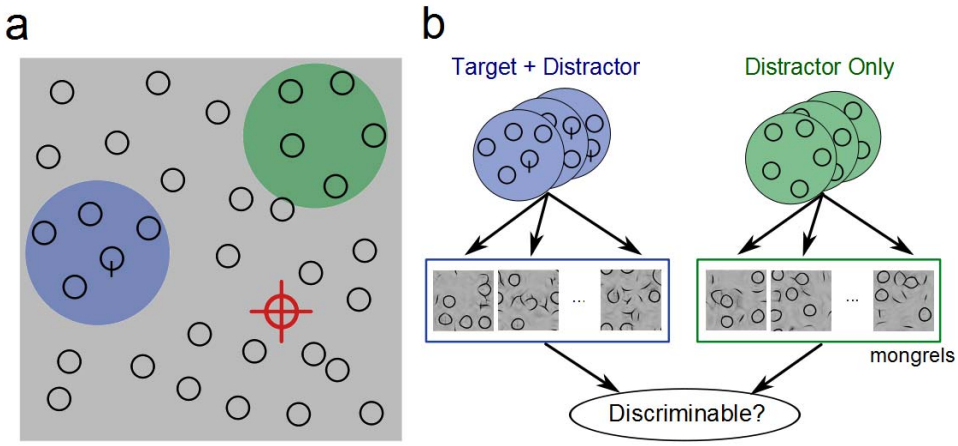


Figure 7: (a) In visual search, we propose that on each fixation (red cross), the visual system computes statistics over a number of local patches. Some of these contain a target and distractors (blue), whereas most contain only distractors (green). The job of the visual system is to distinguish between promising and unpromising peripheral patches and to move the eyes accordingly. (b) We hypothesize, therefore, that peripheral patch discriminability, based on a rich set of summary statistics, critically limits search performance. To test this, we select a number of target + distractor and distractor-only patches, and use texture synthesis routines to generate a number of patches with the same statistics ("mongrels"). We then ask human observers to discriminate between target + distractor and distractor-only synthesized patches, and examine whether this discriminability predicts search difficulty.

Figure 6 shows the results. See Ref. 13 for more details. Performance on the crowding and mongrel classification tasks was significantly correlated (Pearson's $R^2 = 0.65$, $p < 0.01$, one-tailed), and the slope of the regression line (1.2) was not significantly different from 1 ($t(7) = 0.57$, $p>0.20$). This indicates that the summary statistics constrain task performance in a similar way as crowding. Mongrels – and the summary statistics they visualize – capture much of the information maintained and lost under conditions of crowding.

2.3 Visual search

As mentioned in Sec. 1, another part of vision where performance seems limited by an information bottleneck is visual search. Rethinking visual search in light of our model provides immediate insights. If early visual representation is in terms of summary statistics computed over pooling regions that grow with eccentricity, then for typical search displays many of those pooling regions will contain more than a single item. This suggests that rather than thinking about the similarity between a single target and a single distractor, we should be thinking about the similarity between peripheral patches containing a target (plus distractors) and those containing only multiple distractors. According to our model, that is the visual system's real task as it confronts a search display, as illustrated in Figure 7a.

In Figure 7a, the target ('Q') is not visible near the current fixation (red crosshairs), so the subject continues searching. Where to look next? A reasonable strategy is to seek out regions that have promising statistics. The green and blue discs represent two hypothetical pooling regions in the periphery, one containing the target (plus distractors), the other containing only distractors. If the statistics in a target-present patch are noticeably different from those of target-absent patches, then this can guide the subject's eyes toward the target. However, if the statistics are inadequate to make the distinction, then the subject must proceed without guidance.

The simple prediction is that search will be easy if and only if the visual statistics of target-present patches are sufficiently different from those of target-absent patches[†]. Using a methodology similar to that described in Sec. 2.2, we make mongrels of target-present and target-absent patches, and ask how well observers can distinguish between them (Figure 7b), as a measure of the inherent difficulty discriminating based upon the summary statistics.

Figure 8 shows several mongrels for each of 3 conditions: feature search for a tilted line among vertical; conjunction search for a white vertical among white horizontals and black verticals; and configuration search for a T among L's. It is worth examining these mongrels in more detail. Search for a tilted line among vertical is known to be easy.[4] The target-present mongrels for this condition clearly show a target-like item, whereas the distractor-only mongrels do not. Patch discrimination based upon statistics alone should be easy, predicting easy search. The task should be possible in the periphery, without moving one's eyes. Conjunction search for a white vertical among black verticals and white horizontals shows some intriguing "illusory conjunctions"[4,34] – white verticals – in the distractor-only mongrels. This makes the patch discrimination task more difficult, and correctly predicts more difficult search. Search for a 'T' among 'L's is known as a difficult "configuration search".[35] In fact, the mongrels for this condition show 'T'-like items in some of the distractor-only patches, and no 'T'-like items in some of the target+distractor mongrels. Patch discrimination based upon statistics looks difficult, predicting difficult search.

Figure 9 plots search performance for 5 classic search tasks, versus the results of our mongrel discrimination experiment. As is standard in the search literature, we quantify search difficulty as the slope of the function relating mean reaction time to the number of items in the display. The results agree with our predictions. When target+distractor patch statistics are similar to distractor-only statistics, search is slow; when the statistics are dissimilar, search is fast. The data shows a clear relationship between search performance and visual similarity of patch statistics as measured by human discrimination of the mongrels ($R^2=.98$, $p<0.01$).[36] We have also recently demonstrated that, by developing an ideal observer model which chooses the next fixation based upon the information available in peripheral summary statistics, it is possible to quantitatively predict the mean number of fixations required to find the target for these search conditions.[37] Note that any differences between feature search vs. conjunction vs. configuration are captured by the statistical discriminability of target-present patches from target-absent. No additional attentional mechanisms are required to bind features in the conjunction and configuration search conditions but not in feature search.

[†] This assumes all else being equal at later stages of vision. Consider, for example, search for an "N" among mirror-reversed "N"s. In terms of summary statistics, this should be the same as search for a mirror-reversed "N" among "N"s. However, the two conditions might not be equivalent due to more effective processing of the familiar "N"s at later stages. Here we examine how far one can get in explaining search performance based only on the low-level mechanisms modeled here, without invoking higher-level mechanisms.

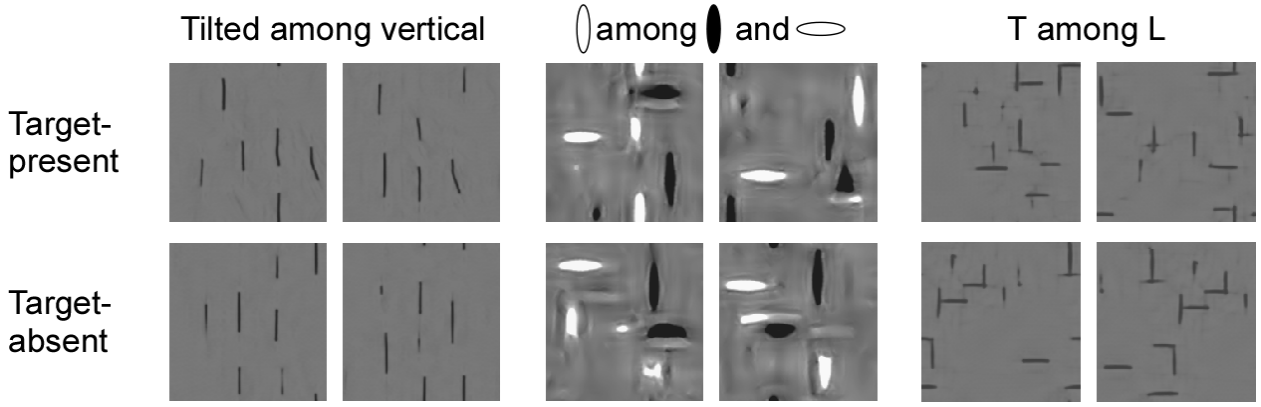


Figure 8. Example mongrels for target-present (row 1) and target-absent (row 2) patches, for 3 classic search conditions: (a) tilted among vertical; (b) orientation-contrast conjunction search; (c) T among L. How discriminable are target-present from target absent mongrels? Inspection suggests that the summary statistic model correctly predicts easy search for tilted among vertical, more difficult conjunction search, and yet more difficult search for T among L, as indicated by experimental results (Fig. 9).

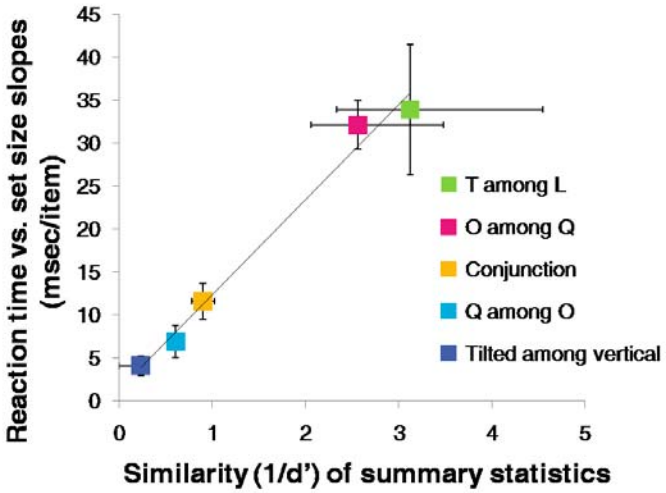


Figure 9. Testing whether the Texture Tiling model predicts visual search performance. The y-axis shows search performance for correct target-present trials, as measured by search efficiency, i.e. the mean number of milliseconds (msec) of search time divided by the number of display items. The x-axis shows a measure of the statistical similarity of target+distractor and distractor-only patches, based on the empirical discriminability, d', of the corresponding mongrels. Clearly there is a strong relationship between visual search difficulty and difficulty distinguishing between target+distractor and distractor-only patches based on their summary statistics, in agreement with our predictions. (y-axis error bars = standard error of the mean; x-axis error bars = 95% confidence intervals for 1/mean(d').)

3. IMPLICATIONS FOR IMAGING APPLICATIONS

As argued in Sec. 1, a critical constraint on task performance is what the user can do with a single glance. Based on the information available in a glance, the user moves their eyes, scanning the display for items of interest, and piecing together a coherent view of the display. At any given instant, much of a display appears where the user is "not looking." We have proposed that where one is "not looking," the visual system represents its inputs by a rich set of summary statistics, in order to get through a bottleneck in visual processing. Previous sections have demonstrated that our model can predict both crowding in peripheral vision[13] and visual search performance.[36,37] Perhaps this model could inform design of better information visualizations and user interfaces, by enabling designers to make displays that effectively guide eye movements and make important information available in a glance. However, it is difficult to attend to our peripheral vision to gain insights about the information available there. And it is difficult to intuit what information is available in a complex set of summary statistics. However, we can use our model to generate visualizations of the information available in a glance, as suggested by Raj & Rosenholtz.[38]

In previous sections we used mongrels to visualize the information available in a single pooling region. To visualize the information available across the visual field in a glance, we need to combine information across pooling regions that tile the visual field (more details of this tiling were given in Sec. 2.1.2). Given a fixation point, we locate a number of overlapping pooling regions on the input image. Each pooling region is sized according to Bouma's law, and within each pooling region we compute the summary statistics described in Section 2.1. Synthesis is initiated by assuming the foveal region (a small circle about fixation) is reconstructed perfectly. Then, moving outward, each subsequent pooling region is synthesized using the previous partial synthesis result as the seed for the texture synthesis process. The process iterates a number of times over the entire image. We use a coarse-to-fine strategy to speed convergence. This procedure is not guaranteed to converge, but it yields plausible results sufficient for gaining intuitions.

Take, for instance, the pair of New York subway maps shown in Fig. 10ab. For most of the history of the New York subway, its maps were fairly geographically accurate, with stops shown approximately at their actual locations, parks and boroughs shown an appropriate size and scale, and so on, as in Fig. 10a. This realism was and continues to be atypical. London tube maps, for instance, have nearly always been abstract, like circuit diagrams more than maps, with few geographical landmarks depicted, and lines drawn at a limited set of angles (0, 90, 45, and 135 degrees).. In 1972, the map switched to a far more schematic map, designed by Massimo Vignelli. (Figure 10b shows a newer, unofficial design from Vignelli (2008), which is completely abstract and contains no geographic features at all.) The Vignelli map was cleaner, less cluttered, and a style preferred in much of the world. New Yorkers hated it for its geographic inaccuracies; graphic designer Michael Bierut has suggested that, due to Manhattan's street grid, New Yorkers have better than average understanding of their city's geography, and thus are more bothered by the errors.[39] Issues with the inaccurate geography are valid but highly cognitive. Visually, can we get confirmation that the Vignelli map is less cluttered, and thus more comprehensible at a glance? If so, this would help explain the preference for more abstract maps in much of the world.

Figure 10c shows a mongrel of Fig. 10a. Near fixation, the map is quite good, and it maintains much of its structure north and south of that point as well. However, beyond that the map is a mess. What lines connect to the east, for instance? Figure 10d is a mongrel of Fig. 10b. It is a nearly perfect replica of the original map. Though lacking in geographic detail, it is uncluttered enough that the information available at a glance is rich and veridical, with an obvious exception of not being able to read the peripheral text. Introspection on the original maps can yield a similar conclusion. However, this sort of visualization should prove valuable because it enables insights without the difficulty attending to and introspecting on peripheral vision.

4. DISCUSSION

We have proposed the Texture Tiling model for early visual system representation. According to this model, the visual system computes local summary statistics over pooling regions that overlap, tile the visual field, and grow linearly with eccentricity. Other researchers have previously theorized that the visual system might compute summary statistics under certain circumstances. For example, Rensink,[40] Treisman,[41] and Wolfe[42] have all sketched systems in which, in the absence of attention, the visual system might encode statistical properties, which might in turn be used to determine such things as the "gist" of a scene. Treisman[41] and Wolfe[43] proposed that statistics might be computed in parallel across the visual field, whereas Rensink[40] suggested that "proto-objects" might be what was computed in parallel, with a limited capacity process computing some sort of statistics over the proto-objects to extract gist and layout. From this point of

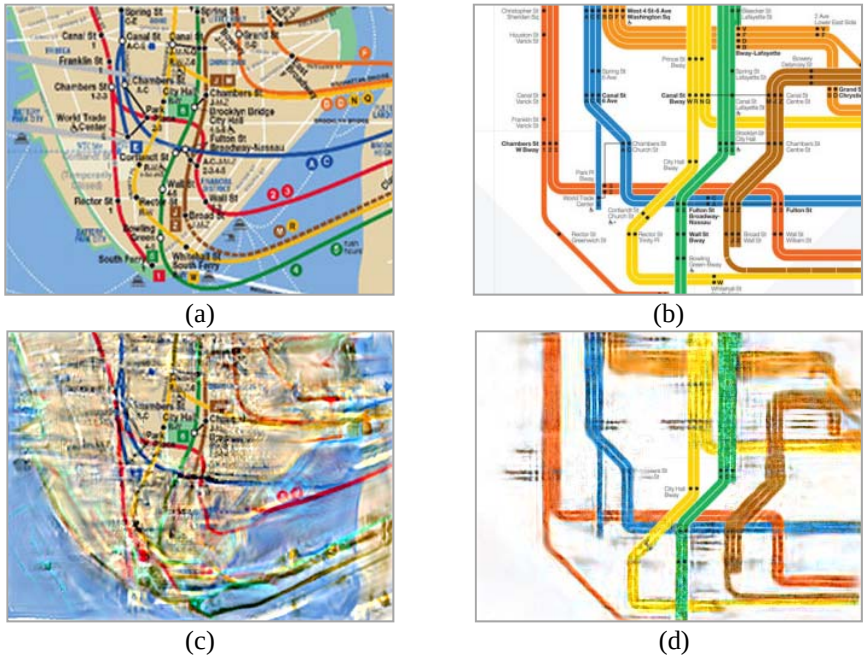


Figure 10. (a) Section of standard New York city subway map. (b) Approximately the same section of an abstract schematic map designed by Vignelli, 2008. When fixating (a) and (b), what information is available? (c) Full-field mongrel of (a), fixation at City Hall (near green patch slightly above center). Near fixation, the information is fairly veridical, but it becomes confusing further away. (d) Full-field mongrel of (b), fixation at City Hall (pair of black dots near the center). Other than unreadable text and some loss of resolution of the different tracks, our model predicts that the information available in a glance is very nearly the information available in the original map.

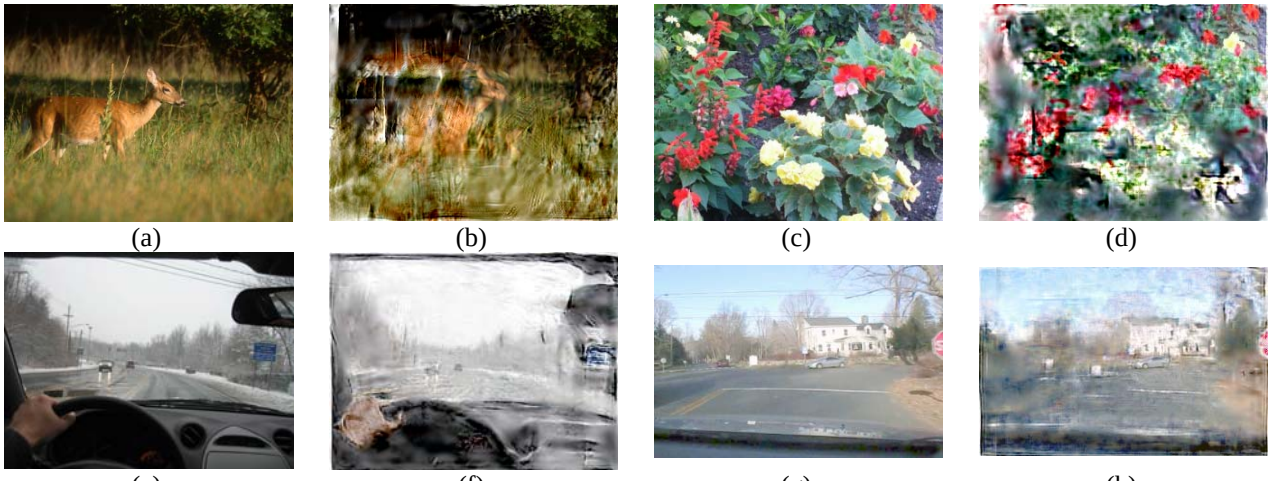


Figure 11. (a),(c) Natural scenes, such as might be used for a task in which observers judge scene gist or animal/no-animal. (b),(d) The corresponding full-field mongrels, assuming fixation at the upper right corner of each original image. The animal/no-animal task should be quite easy from this representation. (e),(g) Scenes out a car windshield. (f),(h) Full-field mongrels, fixation (as typical for driving) on the car ahead. The driver should, without even moving his eyes, be able to follow the road, tell whether the road is wet, determine that he is approaching a curve or intersection, and spot a stop sign. Summary statistics may capture enough information to support "zombie behaviors"[43] like driving on a familiar road.

view, one can think of our model as a testable implementation of some of these earlier ideas: a concrete proposal for the underlying representation.

These previous theories focused on a proposed difference between visual processing with and without selective attention. Treisman[41] and Rensink,[40] suggested that attention is generally required for object perception. Wolfe[42] hypothesized two pathways: one limited capacity and selective, for processing objects, and the other unlimited capacity, for processing gist, computing saliency, extracting statistical summaries, and so on.

Top-down, selective attention clearly has some effect, as indicated by a number of behavioral and physiological studies. One needs some mechanism to account for effects of covert attention and dual-task on visual processing. Consider illusory conjunctions. The Texture Tiling model predicts illusory conjunctions in search due to pooling over too large a region in the periphery. Why would one also perceive illusory conjunctions at the fovea, where pooling regions are small? At the fovea, illusory conjunctions typically occur under dual-task conditions, as in Ref. 34. Clearly this is an effect of attention. Perhaps attending elsewhere makes pooling regions grow somewhat. Arguably, Bouma's law actually specifies the minimum size of the regions *when attending*, as most crowding experiments have been run under conditions of full covert attention. Allowing pooling regions to grow with inattention might be a strategy by which one could extend the Texture Tiling model to account for where one is "not looking" = not attending.

However, attention most likely has a modest effect compared to foveation. Visual crowding indicates that, even with full attention, there are profound, fundamental limits on the information available in the periphery. On the flip side, humans can perform many complex tasks, apparently with neither consciousness nor attention, such as driving home on a familiar route, or reading aloud a children's book. Such "zombie behaviors"[43] would seem to imply that one can remove attention without a huge impact on either visual representation or task performance. Under normal circumstances, representation in early vision might look quite a bit like the model presented in this paper. What tasks can we expect to do with this representation?

There is a loss of information in representing the visual input in terms of summary statistics. This loss of information is what allows the model to correctly predict that some crowded letter recognition tasks will be difficult (Fig. 6). This loss of information also allows us to predict that some visual search tasks will be difficult, while others will be easy (Fig. 9). One might ask whether the loss of information is too severe to allow performance of other visual tasks, such as extracting the "gist" from a scene. Figure 11 gives some intriguing demos. The details of a scene, outside the fovea, are murky: the shape of the deer is unclear, and the grass is badly represented (Fig. 11ab); the exact layout of the flowers is incorrect (Fig. 11cd); there's clearly an oncoming vehicle, but what does it look like (Fig. 11ef)?; and the windows of the house are poorly localized (Fig. 11gh). This murkiness of the details is perhaps to be expected from the phenomenon of change blindness.[2,3] Nonetheless, the model predicts we should be able to perform a number of useful tasks with a summary statistic representation. Even in the periphery we would expect an animal/no-animal task to be fairly easy. Even if one cannot clearly identify the deer or its pose, we are fairly sure it stands outdoors, in some grass, and has long ears and big eyes. The no-animal scene almost certainly contains flowering plants. (See Fig. 11a-d.) Even in a glance, a driver should be able to follow the road, detect slippery road conditions, spot oncoming cars, and notice salient signs. This information is likely sufficient to support the "zombie behavior" of driving without paying attention. Both confusion over details and the sense of a rich representation of the world are predicted by our model.

We have proposed that, early on, the visual system compresses its inputs in order to get through the information bottleneck. In the weak version of our hypothesis, other major constraints apply later in visual pathways and help determine performance at visual tasks. In strong version of the hypothesis, representation in terms of summary statistics may be the key constraint in vision. After lossy compression, designed to get as much useful information as possible through a bottleneck, perhaps our visual systems can do whatever tasks that information affords. If the remaining information is sufficient for distinguishing between two possible scene categories, or to identify an object, an observer can do those tasks. If the information is insufficient to find a path through a map in a glance, the user will need to move their eyes. This is an ambitious hypothesis, but as we have demonstrated here, it is rendered testable by advances in methodology that allow one to visualize the information available in a complex set of summary statistics.

ACKNOWLEDGMENTS

Thanks to Alvin Raj for help with the demonstrations, and to Alvin, Benjamin Balas, Ted Adelson, Jie Huang, and Phillip Isola for useful discussions. This work was supported by grants to Dr. Rosenholtz from the NSF (BCS-0518157), NIH (1-R21-EU-10366-01A1), and Qualcomm.

REFERENCES

1. J. Hochberg, "In the mind's eye," *Annual Meeting, American Psychological Association, 1966*, R. N. Haber, ed., *Contemporary Theory and Research In Visual Perception*, pp. 309-331, New York, Holt, Rinehart, and Winston, 1968.
2. R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science*, **8**, pp. 368-373, 1997.
3. D. J. Simons and D. T. Levin, "Change blindness," *Trends in Cognitive Sciences*, **1**, pp. 261-267, 1997.
4. A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, **12**, pp. 97-136, 1980.
5. J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the Feature Integration model for visual search," *J. Experimental Psychology: Human Perception & Performance*, **15**, pp. 419-433, 1989.
6. M. C. Potter, "Short-term conceptual memory for pictures," *J. Experimental Psychology: Human Learning and Memory*, **2**, pp. 502-522, 1976.
7. B. Pinna and R. L. Gregory, "Shifts of edges and deformations of patterns," *Perception*, **31**, pp. 1503-1508, 2002.
8. S. R. Ellis and L. Stark, "Eye movements during the viewing of Necker cubes," *Perception*, **7**, pp. 575-581, 1978.
9. N. Kawabata, K. Yamagami, and M. Noaki, "Visual fixation points and depth perception," *Vision Research*, **18**, pp. 853-854, 1978.
10. G. C. –W. Shyi and M. A. Peterson, "Perceptual organization in a brief glance: The effects of figure size, figure location, and the attentional focus," *Chinese Journal of Psychology*, **34**, pp. 1-18, 1992.
11. R. A. Rensink, "Change detection," *Annual Review of Psychology*, **53**, pp. 245-277, 2002.
12. R. Van Rullen, L. Reddy, and C. Koch, "Visual search and dual tasks reveal two distinct attentional resources," *J. Cognitive Neuroscience*, **16**, pp. 4-14, 2004.
13. B. Balas, L. Nakano, and R. Rosenholtz, "A summary-statistic representation in peripheral vision explains visual crowding," *Journal of Vision*, **9**, pp. 1-18, 2009. http://journalofvision.org/9/12/13.
14. J. Portilla and E. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, **40**, pp. 49-71, 2000.
15. B. Balas, "Texture synthesis and perception: Using computational models to study texture representations in the human visual system," *Vision Research*, **46**, pp. 299-309, 2006.
16. L. Parkes, J. Lund, A. Angelucci, J. A. Solomon, and M. Morgan, "Compulsory averaging of crowded orientation signals in human vision," *Nature Neuroscience*, **4**, pp. 739-744, 2001.
17. D. M. Levi, "Crowding – an essential bottleneck for object recognition: A mini-review," *Vision Research*, **48**, pp. 635-654, 2008.
18. D. G. Pelli, and K. A. Tillman, "The uncrowded window of object recognition," *Nature Neuroscience*, **11**, pp. 1129-1135, 2008.
19. B. Julesz, "A theory of preattentive texture discrimination based on the first order statistics of textons," *Biological Cybernetics*, **41**, pp. 131-138, 1981.
20. J. Beck, "Textural segmentation, second-order statistics, & textural elements," *Biological Cybernetics*, **48**, pp. 125-130, 1983.
21. H. Voorhees and T. Poggio, "Computing texture boundaries from images," *Nature*, **333**, pp. 364-367, 1988.
22. R. Rosenholtz, "Significantly different textures: A computational model of pre-attentive texture segmentation," *European Conference on Computer Vision*, D. Vernon, ed., *LNCS*, **1843**, pp. 197-211, Springer, Berlin, 2000.
23. D. R. T. Keeble, F. A. A. Kingdom, B. Moulden, and M. J. Morgan, "Detection of orientationally multimodal textures," *Vision Research*, **14**, pp. 1991-2005, 1995.
24. G. Van de Wouwer, P. Scheunders, and D. Van Dyck, "Statistical texture characterization from discrete wavelet representations," *IEEE Transactions on Image Processing*, **8**, pp. 592-598, 1999.
25. R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," *Vision Research*, **39**, pp. 3157-3163, 1999.
26. I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson, "Image statistics and the perception of surface qualities," *Nature*, **447**, pp. 206-209, 2007.
27. D. Ariely, "Seeing sets: Representation by statistical properties," *Psychological Science*, **12**, pp. 157-162, 2001.
28. S. C. Chong and A. Treisman, "Representation of statistical properties," *Vision Research*, **43**, pp. 393-404, 2003.
29. S. C. Chong and A. Treisman, "Statistical processing: Computing the average size in perceptual groups," *Vision Research*, **45**, pp. 891-900, 2005.

30. K. Fukushima, “Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, **36**, pp. 193–202, 1980.
31. M. Riesenhuber and T. Poggio, “Hierarchical models of object recognition in cortex,” *Nature Neuroscience*, **2**, pp. 1019-1025, 1999.
32. J. Freeman and E. Simoncelli, “Crowding and metamerism in the ventral stream,” *Vision Sciences Society Annual Meeting*, abstract in *Journal of Vision*, **10**, p. 1347, 2010.
33. A. J. Bell and T. J. Sejnowski, “The ‘independent components’ of natural scenes are edge filters,” *Vision Research*, **37**, pp. 3327-3338, 1997.
34. A. Treisman and H. Schmidt, “Illusory conjunctions and the perception of objects,” *Cognitive Psychology*, **14**, pp. 107-141, 1982.
35. J. M. Wolfe, K. R. Cave, and S. L. Franzel, “Guided search: An alternative to the Feature Integration model for visual search,” *J. Experimental Psychology: Human Perception & Performance*, **15**, pp. 419-433, 1989.
36. R. Rosenholtz, S. Chan, and B. Balas, “A crowded model of visual search,” *Vision Sciences Society Annual Meeting*, abstract in *Journal of Vision*, **9**, p. 1197, 2009. Paper in preparation.
37. R. Rosenholtz, L. Ilie, and B. J. Balas, “An ideal saccadic targeting model acting on pooled summary statistics predicts visual search performance,” *Vision Sciences Society Annual Meeting*, abstract in *Journal of Vision*, **10**, p. 1293, 2010. Paper in preparation.
38. A. Raj and R. Rosenholtz, “What your design looks like to peripheral vision,” *Proc. 7th Symposium on Applied Perception in Graphics and Vision* (APGV ’10), pp. 89-92, ACM, New York, 2010.
39. A. Mindlin, “Win, lose, or draw: The great subway map wars,” *The New York Times*, Sept. 3, 2006, http://www.nytimes.com/2006/09/03/nyregion/thecity/03maps.html.
40. R. A. Rensink, “Change blindness: Implications for the nature of visual attention,” in *Vision and Attention*, M. Jenkin and L. Harris, eds., pp. 169-188, Springer, New York, 2001.
41. A. Treisman, “How the deployment of attention determines what we see,” *Visual Cognition*, **14**, pp. 411-443, 2006.
42. J. M. Wolfe, “Guided Search 4.0: Current Progress with a model of visual search,” in W. Gray ed., *Integrated Models of Cognitive Systems*, pp. 99-119, Oxford, New York, 2007.
43. C. Koch and F. Crick, “The zombie within,” *Nature*, **411**, p. 893, 2001.