A summary-statistic representation in peripheral vision explains visual crowding

Benjamin Balas

Laboratories of Cognitive Neuroscience,
Children's Hospital Boston, Boston, MA, USA

Lisa Nakano

Department of Brain and Cognitive Sciences, MIT,
Cambridge, MA, USA

Ruth Rosenholtz

Department of Brain and Cognitive Sciences, MIT,
Cambridge, MA, USA

Peripheral vision provides a less faithful representation of the visual input than foveal vision. Nonetheless, we can gain a lot of information about the world from our peripheral vision, for example in order to plan eye movements. The phenomenon of crowding shows that the reduction of information available in the periphery is not merely the result of reduced resolution. Crowding refers to visual phenomena in which identification of a target stimulus is significantly impaired by the presence of nearby stimuli, or flankers. What information is available in the periphery? We propose that the visual system locally represents peripheral stimuli by the joint statistics of responses of cells sensitive to different position, phase, orientation, and scale. This "textural" representation by summary statistics predicts the subjective "jumble" of features often associated with crowding. We show that the difficulty of performing an identification task within a single pooling region using this representation of the stimuli is correlated with peripheral identification performance under conditions of crowding. Furthermore, for a simple stimulus with no flankers, this representation can be adequate to specify the stimulus with some position invariance. This provides evidence that a unified neuronal mechanism may underlie peripheral vision, ordinary pattern recognition in central vision, and texture perception. A key component of our methodology involves creating visualizations of the information available in the summary statistics of a stimulus. We call these visualizations "mongrels" and show that they are highly useful in examining how the early visual system represents the visual input. Mongrels enable one to study the "equivalence classes" of our model, i.e., the sets of stimuli that map to the same representation according to the model.

Introduction

Evidence from visual search, from change blindness, and from dual task experiments indicates that there is a bottleneck in visual processing (Nakayama, 1990). Our visual systems seem to represent foveal, attended stimuli with a fair degree of fidelity, but more coarsely encode stimuli with increasing eccentricity, and/or when those stimuli are unattended. The vast majority of vision is represented with this lower fidelity. Understanding the information available to the visual system in the coarse encoding is key to understanding our visual capabilities and limitations. For example, if a task can be done using peripheral vision, e.g., detecting a target that "pops out," then that task should be relatively easy, since it will not require moving one's eyes to bring new parts of the stimulus into the fovea. At the other extreme, if the information available peripherally is uninformative for an observer's task, the task will be difficult, as it will require

moving the eyes over the image, and little information may be available to help guide these eye movements.

In spite of the importance of peripheral vision, there is little understanding of the information available to the visual system, or of the visual representation, in peripheral vision. Peripheral vision has mostly been characterized in terms of the reductions in resolution or contrast sensitivity as eccentricity increases (Anstis, 1974; Peli & Geri, 1999; van Essen & Anderson, 1990; Virsu & Rovamo, 1979). An important exception has been the study of visual crowding. Visual crowding refers to the phenomenon in which a target may be easily recognizable when viewed in isolation but becomes difficult to identify when flanked by other items. The ease of recognizing the isolated target indicates that crowding is not simply a by-product of reduced visual acuity in the periphery. Instead, it seems that the visual system applies some as-yet-unspecified lossy transformation—perhaps some form of "feature integration," pooling, or averaging (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Pelli, Palomares, & Majaj,

2004)—to the stimulus, resulting in the subjective experience of mixed-up, jumbled visual features (Martelli, Majaj, & Pelli, 2005).

In this paper, we will use a novel experimental paradigm to test the hypothesis that the increasing loss of information with increasing eccentricity is due to a representation of the visual input in terms of summary statistics. By "summary statistics" we mean that the visual input is represented, locally, by measurements that summarize that local region.[1] Specifically we hypothesize that over some local pooling region, the visual system computes joint statistics of responses of oriented V1-like feature detectors across different orientations, spatial locations, and so on. Later sections give details of the particular statistical representation tested. This encoding will, for instance, make it simple to determine if a region contains some vertically oriented structure but does not allow one to say whether there was a strong vertical orientation at a particular location (except in trivial cases, e.g., in which the entire region is filled with strong vertical structure).

Representing the visual stimuli in this manner involves a substantial loss of information. Thus, one should look for evidence of this representation in situations in which visual performance is compromised. Visual crowding provides an example. One can experience the phenomenon of crowding by fixating in the center of Figure 1a. In the left periphery, one can easily read the isolated letter "G". In the right periphery, the crowding caused by the flanking letters makes the same letter hard to read. A simple reduction in resolution cannot make the G on the right illegible while preserving the legibility of the G on the left. Indeed, subjectively the difficulty reading the crowded G does not stem from a too-blurry stimulus. The percept of the crowded stimulus contains sharp, letter-like forms, but the exact details seem lost in a jumble, as if each letter's features (e.g., vertical bars and rounded curves) have come untethered from each other and been incorrectly bound to the features of neighboring letters. Indeed, crowding is sometimes described as a failure of binding (Pelli et al., 2004). Crowding also occurs with stimuli other than letters (Andriessen & Bouma, 1976; Levi, Hariharan, & Klein, 2002; Martelli et al., 2005; Parkes et al., 2001; van den Berg, Roerdink, & Cornelissen, 2007; Wilkinson, Wilson, & Ellemberg, 1997). The phenomenon is not restricted to the periphery but is more noticeable in peripheral vision, where even with a fair amount of space between the target and its neighbors, the neighbors still impair identification performance.

In addition to providing a phenomenon in which the visual system seems "broken," investigating crowding allows one to draw upon extensive theoretical and empirical advances made in recent years to functionally differentiate crowding from related phenomena (such as lateral masking), and to understand the multiple stimulus factors which modulate the size of crowding effects. Presently, crowding has been demonstrated with a number of tasks, including letter identification (Flom, 1991; Pelli et al., 2004), face recognition (Martelli et al., 2005), and the discrimination of orientation (Levi et al., 2002; Parkes et al., 2001; Wilkinson et al., 1997), contrast, spatial frequency, size, saturation, and hue (Andriessen & Bouma, 1976; van den Berg et al., 2007). The "critical spacing" between stimuli necessary to achieve crowding is relatively independent of the contrast of the flanking elements and of target size (Levi et al., 2002; Strasburger, Harvey, & Rentschler, 1991). This invariance to stimulus contrast and size suggests that the critical spacing may reflect an intrinsic limitation imposed on the visual system either by the architecture of neural pooling across peripheral vision, or the resolution of attention. There is also a more or less equal effect of crowding over a range of flanker size (10:1), flanker number (Pelli et al., 2004; ≥2), and across certain variations in flanker type (letter, black disk, or square; Eriksen & Hoffman, 1973; Loomis, 1978). However, many studies have documented systematic effects of the similarity of target and flanker (Chung, Levi, & Legge, 2001; Donk, 1999; Estes, 1982; Ivry & Prinzmetal, 1991; Kooi, Toet, Tripathy, & Levi, 1994; Nazir, 1992).

Despite widespread interest in crowding as a basic visual phenomenon and as a methodological tool for investigating broader issues in visual neuroscience, there is as yet no explicit computational theory of what representations or computations in the visual system lead to crowding. Our efforts here thus provide an important complement to the existing crowding literature. A great deal of empirical data can be summarized by what has been called the "Bouma law"—that the "critical spacing" between items that determines whether or not crowding occurs is roughly half the eccentricity (Pelli & Tillman, 2008). However, we still lack a predictive model that can answer basic questions about crowded perception. Given an arbitrary stimulus, we currently have no models that can predict what information an observer will be able to obtain from the display. We cannot say what tasks will be easy and what tasks will be hard.

The "jumbled" appearance of crowded stimuli has recently led some researchers (Parkes et al., 2001; Pelli et al., 2004) to suggest that in a crowded display perception might be more like texture perception than like fully attentive object perception. The exact meaning of "texture perception" has been left unclear, but the implication is that the information kept consists of summary statistics, comprising information about distributions of feature values rather than localized feature

G   +   K G P

Figure 1. The phenomenon of crowding. The letter "G" on the left is clearly identifiable in the periphery, when fixating on the "+". However, the presence of flanking letters makes recognition of the "G" on the right quite difficult.

maps. Such statistical properties likely mediate texture segmentation and discrimination (Balas, 2006; Beck, 1983; Julesz, 1981; Keeble, Kingdom, Moulden, & Morgan, 1995; Rosenholtz, 2000; Voorhees & Poggio, 1988), and visual saliency (Rosenholtz, 1999, 2001) and also contain important information for material perception (Motoyoshi, Nishida, Sharan, & Adelson, 2007) and scene perception (Greene & Oliva, 2009; Oliva & Torralba, 2001; Walker-Renninger & Malik, 2004).

For one particular task and set of stimuli, the notion that computation of summary statistics underlies crowding has been formalized to the point of testing quantitative predictions: Parkes et al. (2001) have shown that for displays of oriented Gabors, performance at estimating whether target lines are tilted to the left or right of vertical is well predicted by a model that bases its decisions upon the average orientation of the Gabors. This model constitutes an important theoretical advance, in that it generates clear quantitative predictions about a particular class of crowded stimuli. However, it is important to note several limitations of this modeling effort.

First, the jury is still out as to whether the visual system is *forced* to compute summary statistics in the periphery. Parkes et al.'s displays contained more than one tilted target, and location was not a cue to whether a given item was a target. Computing average orientation might simply have been a better strategy than judging target tilt based upon the item that appeared most tilted.

Furthermore, Parkes et al.'s model is very limited in scope. It is not obviously applicable to more complicated displays, or to peripheral vision more broadly. Confronted with a task of identifying a single letter in an array of letters, how do we apply Parkes et al.'s model to generate a prediction? There are several issues, here, all of which we will address in this paper. (1) Clearly other statistics are necessary, and we need a hypothesis as to which ones. (2) We need to be able to measure those statistics in an arbitrary image. This will facilitate testing on more complex imagery, a requirement for examining a more complete set of hypothesized statistics. The scope of a model is extremely limited if it requires the experimenter to "tell it" what features the experimenter *intended* to put into the display—e.g., the orientations of the Gabors. Models that cannot make measurements on actual images will fail if there are emergent features that the experimenter does not tell the model about, e.g., alignment of neighboring Gabors. Models that operate on images as input make testable predictions about how performance will change if one adds noise or blur to a stimulus, changes the contrast, and so on. (3) More complete sets of statistics and more complex experimental stimuli necessitate a new methodology for testing whether these models capture the information available in peripheral vision.

Our goal is to build upon Parkes et al.'s proposal that ensemble properties are encoded in peripheral vision, expanding upon the set of statistical features under consideration, and widening the scope of our model to encompass a much wider range of inputs. The resulting model will give us insight into the puzzling phenomenon of crowding. Crowding is often treated as an oddity, a failure that is to be measured but not explained. We will argue that crowding is a part of ordinary vision that emerges naturally from constraints imposed by goals of the visual system and the bottleneck of vision.

In what follows, we first describe our methodology, which can be applied to testing a broad class of models. We will make use of recent work in parametric texture analysis/synthesis techniques from computer graphics. This work suggests a candidate set of statistics that the visual system might compute in peripheral vision. The proposed computations reflect known aspects of early visual processing. Furthermore, texture analysis/synthesis techniques enable us to *visualize* the information available in a given set of statistics. We will use these visualizations to test our model.

A methodology for testing models of visual representation

Our methodology has two key components. The first component allows us to visualize the information encoded in a given representation. For this, we use a technique from computer graphics: parametric texture analysis/synthesis (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000). The goal of texture analysis/synthesis routines is to create a new patch of texture that appears to have the same texture as a sample patch. The analysis routines operate by measuring some set of statistics in the input patch. Then, synthesis begins with an arbitrary image-often a sample of random noise—and iteratively applies constraints derived from the measured statistics. The result, after a number of iterations, is a new image that has (approximately) the same statistics as the original patch.

An important insight is that we can apply these techniques to arbitrary images, not just to textures (nor to just classical psychophysical stimuli). Many useful statistics can be measured in arbitrary images; the aforementioned texture analysis/synthesis methods extract statistics such as the marginal and joint statistics of responses of multiscale oriented filters (wavelets). Thus, for arbitrary stimuli, we can synthesize new image "samples" that have the same statistics as the original stimulus. We call these texture synthesized versions of stimuli, "mongrels." ("Mongrel" derives from the Old English "gemong," meaning "crowd". Its meaning as a cross between different types of things ("mongrel", Merriam-Webster Online Dictionary, 2008) evokes the typical crowded percept of a mixture of features from the different items present). An example of a mongrel generated from a simple crowding demonstration is presented in Figure 2. Here we used the statistics computed by Portilla and Simoncelli (2000).
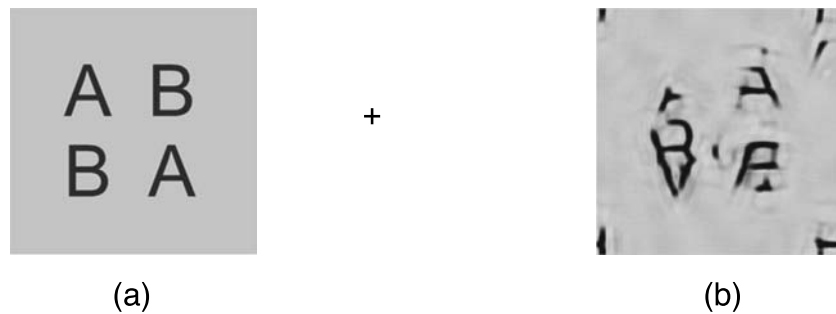
Figure 2. (a) Crowded, when fixating on the "+". (b) A sample mongrel for (a). This visualization of the crowded percept shows the expected mixing of features while preserving sharp edges and high contrast.

Because mongrels share a predefined set of statistics with the original stimulus, a collection of mongrels for a given stimulus can be thought of as a visualization of the information encoded by that set of statistics for that stimulus. The second key component in our methodology consists of using these visualizations to test whether the hypothesized representation underlies peripheral perception.

Consider the analogy of testing a model of dichromat vision. Dichromats have impaired color vision, in which one of the three basic color vision mechanisms performs poorly or is absent entirely. This results in a reduction in the available visual information and leads to poor performance at certain visual tasks, as evidenced by difficulty identifying symbols in Ishihara test plates (designed to test for dichromacy). Predicting task difficulty for dichromats is important both for basic science and so that displays can be designed so they are not confusing to color blind individuals. To this end, researchers have attempted to visualize perception under dichromacy (Brettel, Viénot, & Mollon, 1997). For a

given input image, simulation procedures like these generate a new image, which is intended to give normal observers a sense of what it is like to be a dichromat viewing the input image. An example is shown within Figure 3.

How does one test whether visualizations of dichromacy are any good? Each colored patch in the original stimulus will correspond to a patch in the simulated dichromat's view. However, the key question is not whether the appearance of the simulated patch to a normal observer exactly matches the dichromat percept of the original. Answering this question is philosophically questionable; independent of dichromat simulations, we can never know if one person's percept of a given color is the same as another person's. Furthermore, in judging whether a display will be usable by a dichromat, the issue is whether this information loss will make text too hard to read, or make key features insufficiently salient. A useful model of dichromatic vision needs to tell us what can be discriminated and what cannot but does not need to get the
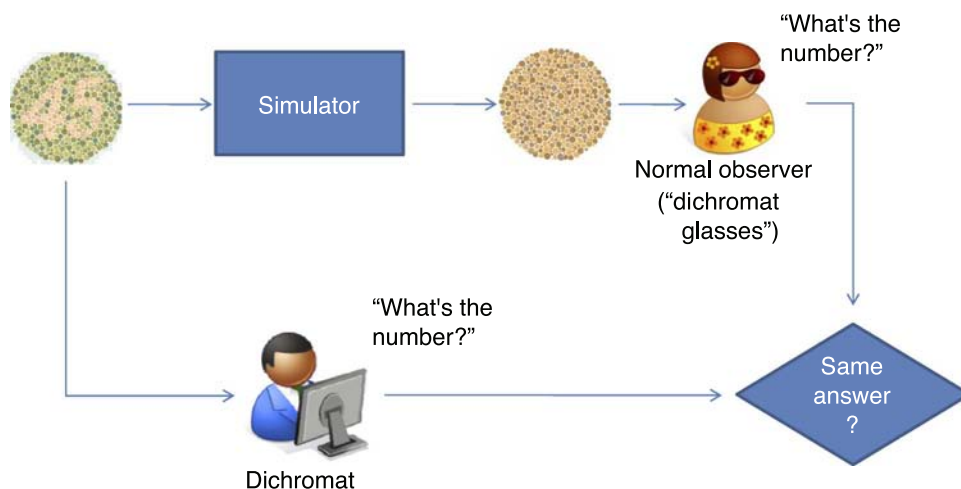

Figure 3. A flowchart representation of a method for verifying a model of dichromacy. A true dichromat performs a task with a given stimulus set, while a normal observer performs that task with stimuli manipulated to reflect what information is lost in the model. If performance of the observers agrees, the model is a useful characterization of dichromacy.

absolute quality of the percept correct. We argue that the key issue is whether the reduction of visual information in the simulated percept matches the reduction in information for the dichromat.

To test whether a dichromat simulation adequately captures the loss of information for a dichromat, one can first generate a number of stimuli (e.g., the Ishihara test plates) and show them to a dichromat. The dichromat then does some task with those stimuli, e.g., identifies the number shown. Next, for each stimulus, one generates a simulation of the dichromat percept. The question, then, is whether the task performance achievable with the *simulations* is predictive of performance by the dichromat with the *original stimuli*. If this is true, over a wide range of stimuli and tasks, then the information captured by the simulations must be roughly the same as the information available to the dichromat, and the simulation procedure is a good one, at least in a functional sense.

The difficulty, here, is in achieving the best task performance possible with the simulations. One could ask a computer vision routine to do the same identification task as the dichromat. However, then our judgment of the quality of our simulations would be highly dependent upon the quality of our computer vision routine. Humans are still far better at identification tasks, in general, than any computer vision routine. Rather than asking a computer vision algorithm to perform the identification tasks, it makes more sense to use the normal human vision system as the recognition system for the identification task since it will outperform any machine vision routines we could create. One would simply show the dichromat simulations to a normal, non-color blind individual, and ask them to do the same task as the dichromat did with the original images. It would be as if the normal observer were doing the task while wearing "dichromat glasses." If, over a wide range of conditions, performance of the normal human observer with the simulated images is predictive of the performance of the impaired dichromat with the original stimuli, then the dichromat simulation must have captured information functionally the same as that available under dichromacy (Figure 3).

We can easily apply this same reasoning and methodology to testing our hypothesized model of peripheral vision. First, the dichromat case posed fundamental difficulties for testing whether a normal percept of a simulation exactly matched the percept of a dichromat. Is this also the case for testing our model—could we not ask whether our mongrels are what people see in peripheral vision? It is less immediately clear that this poses a problem; the peripheral visual system looking at the original displays and the foveal visual system looking at our mongrels would at least belong to the same human being. However, we believe there may be a similar issue. A representation based upon summary statistics is not equivalent to any single image. Much of the time we may make inferences about the world, and choose our next fixation, using only the computed summary statistics, and

with no intervening "image" for the homunculus to view. In somewhat unnatural settings, when we try to attend to a location in the periphery, our visual system likely does its best to comply and give us the sense of an image-like percept. The visual system may well serve up a *number* of "images," which are samples from a process with the measured statistics; akin to our mongrels. This may be the explanation for the shifting and dynamic percept many observers experience when attending to their peripheral vision. Mongrels are visualizations of the information encoded by the summary statistics of a stimulus, and attractive for capturing some subjective characteristics of the peripheral percept, but are not intended as predictions of the percept, per se. As in the dichromat case, we are on more solid ground if we instead attempt to test whether the *tasks* that one can do in peripheral vision match the *tasks* one can do with the mongrel simulation of the available information. In addition, one can argue, again, that this is in fact the most important aspect of behavior for a model to capture.

In terms of the logic presented in Figure 3, in place of the dichromat we will have "impaired observers" whose impairment comes from viewing stimuli peripherally, under conditions of crowding. Performance should be systematically degraded due to a loss of information. For each stimulus used in the crowding task, we will generate a number of mongrels. We will then ask "normal observers" to view these mongrels foveally, for unlimited time, and with the full object and pattern recognition machinery of their visual system. Essentially, it will be like they are viewing the original stimuli with "statistical glasses" (Figure 4)—which induce the same loss of information as the hypothesized representation. As in the dichromat case, the idea is that the normal observers stand in as the best pattern recognizers we can find. By this, we mean that observers are free to use the full capabilities of foveal vision to extract information from the mongrelized image. The assumption is therefore that any losses in performance by these observers are due to the loss of information in the mongrelization process and not to the limits on time, attention, or pattern recognition ability. If, over a wide range of conditions, performance of normal human observers with the mongrels is predictive of the performance of the "impaired" observers viewing the original stimuli in the periphery, then the information encoded by the summary statistics, and visualized by our mongrels, must be the same as the information as that available under peripheral vision. (Note that this need not mean that the *representation* of the information is the same, i.e., there need not be a one-to-one correspondence between neurons and measured statistics.)

This is exactly the approach we have pursued in the current study. We measured performance both under classic crowding conditions and when viewing our mongrels. In both experimental scenarios observers were asked to carry out a 4AFC judgment of letter identity, with a range of flanking stimuli used to modulate performance.
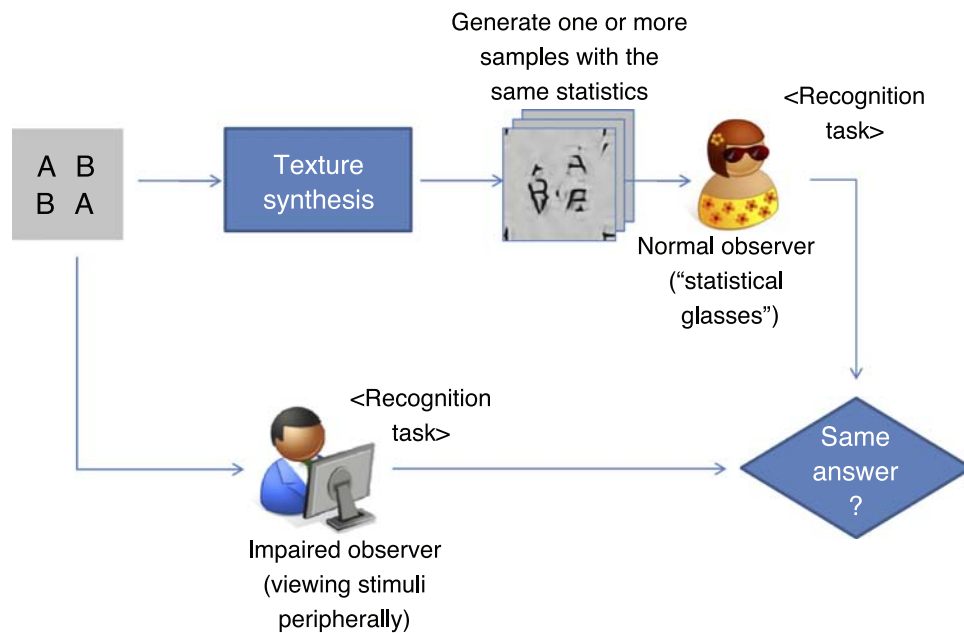
Figure 4. We test the efficacy of our statistical model of peripheral vision by the same methodology outlined in Figure 3. One set of observers views "crowded" arrays peripherally and performs a 4AFC letter recognition task (see text for details). Others do the same task while viewing "mongrels." Critically, these observers foveate the stimuli for unlimited time, making the loss of information due to a statistical representation the prime limiting factor on their performance.

At present, the most reliable way to test what tasks can be done with a certain amount of visual information is a normal human observer. However, it should ultimately also be possible and desirable to replace this human observer with machine classification. The machine classifier would operate directly on the vector of summary statistics extracted from the original stimuli. Replacing the human observer by a machine classifier would ultimately allow us to automatically make predictions without a human observer "in the loop." At present, such a machine classifier is inherently unreliable—we are operating in an approximately 1000-dimensional feature space, have noidea what kind of classifier is appropriate or how to weigh the various dimensions. Much more data is required to answer these questions. However, as an indication of the ultimate feasibility of this approach, we also attempt to predict performance on our crowding tasks via machine classification of statistical vectors derived from the original stimuli, using some simple assumptions for the unknown machine classification parameters. This is described in more detail, below, in the Pattern classification section.

Our hypothesized visual representation

For the purposes of this paper, we assume that within each pooling region, the visual systems collect the same statistics as those used by Portilla–Simoncelli for texture synthesis. This model first measures the responses of V1-like oriented feature detectors. Next, the model computes joint statistics of these features to capture intermediate-level image structure. Though we are not wedded to this particular set of statistics, we chose them both because their model is one of the most powerful parametric texture synthesis models to date, and because Balas (2006) has assessed the perceptual validity of its constituent features in a parafoveal texture discrimination task. The statistics used in this model fall into four main categories: (1) the marginal distribution of luminance in the image; (2) the luminance autocorrelation (capturing the periodicity of the stimulus); (3) correlations of the magnitude of responses of oriented wavelets across differences in orientation, neighboring positions, and across scale; and (4) phase correlation across scale. We describe these statistics in more detail in Appendix A. Figure 5 presents examples of alternative statistical representations in comparison to Portilla–Simoncelli to illustrate the differences in the information captured by these distinct models.[2]

Part of what such a model needs to specify, of course, is the region of the image over which statistics are computed. Work on visual crowding, as described above, suggests that the pooling regions for computing statistics are smallest at the fovea and increase in size approximately linearly with increasing eccentricity. Presumably overlapping pooling regions of this sort tile the entire visual input. For simplicity, we have designed our experiments to place the entire stimulus array, viewed at
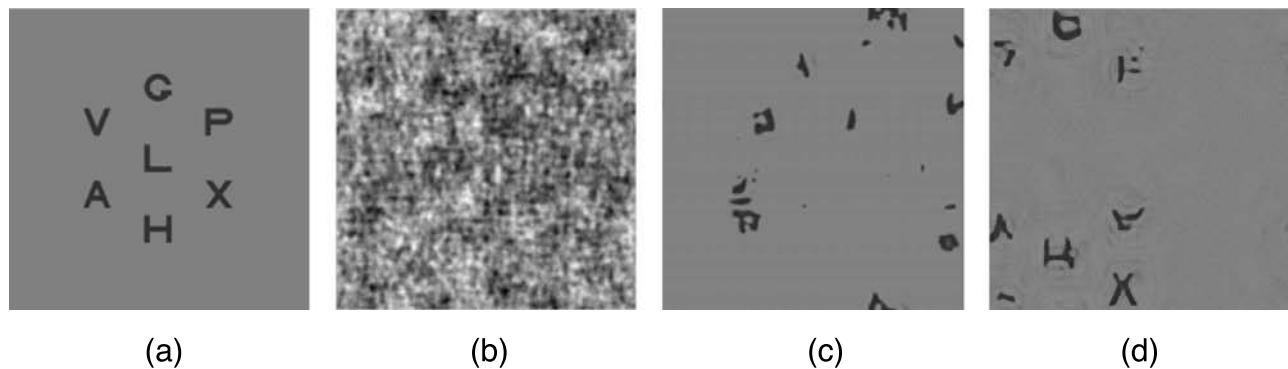
Figure 5. Visualization of the information captured by different choices of summary statistics. (a) Original stimulus. Panels (b)–(d) use a pooling region that covers the entire stimulus. (b) The power spectrum of (a), with random phase. These global statistics poorly capture local structure. (c) Marginal statistics of both the responses of V1-like oriented filters and of intensity values, as in Heeger and Bergen (1995). (d) Marginal distribution of the luminance, joint statistics of responses of V1-like oriented filters, luminance autocorrelation, and phase correlation across scale, as in Portilla and Simoncelli (2000). Both (c) and (d) capture some of the local structure in the original, with (d) capturing more extended structures. We initialized both (c) and (d) with a random seed. These examples do not include a blur to account for peripheral loss of acuity.

the given eccentricity, within a single pooling region. We take the pooling region to be given by the critical spacing specified by the Bouma law.

Experiment: Can a representation based upon summary statistics predict performance at peripheral tasks?

Crowding tasks
Subjects

We tested 3 subjects in our battery of crowding tasks, one author (LN) and two experienced psychophysical observers who were naive to the design and purpose of the experiments. All subjects reported normal or corrected-to-normal vision. At the time of the experiment, LN was naive as to the results of the machine classification and sorting experiments.

Stimuli and procedure

We ran a number of different crowding tasks (Figure 6) with the goal of achieving a wide range of performance across the full set of conditions. In all tasks, the targets were Sloan letters surrounded by a hexagonal array of distracters. Depending on the task, distracters included flanking bars or arcs, additional letters, or grayscale images of natural objects. Participants fixated on a small cross while arrays were presented 14 degrees to the right of fixation for 250 ms. Targets subtended approximately 1 degree in width (except in the large-letter condition,

which contained larger targets and distracters that subtended approximately 3 degrees) and the distance between the center of each target and the center of each distracter was approximately 3 degrees. In all tasks, participants performed a 4AFC recognition task with pre-defined target letters. There were 80 trials per condition, evenly split between the 4 possible targets. Eye movements were monitored online by the experimenter to ensure proper fixation.

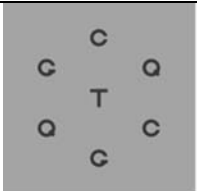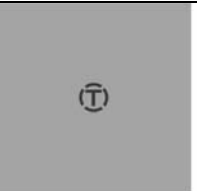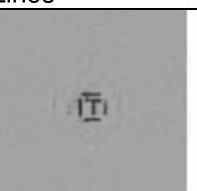### Mongrel sorting tasks: 4AFC letter identification
Subjects
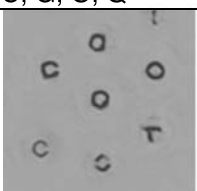
Our sorting tasks were carried out by 6 naive volunteers from the MIT community. All subjects reported normal or corrected-to-normal vision.

Stimuli and procedure

We used the Portilla–Simoncelli (P–S) texture analysis/synthesis algorithm to generate mongrels from the crowded stimuli. We blurred each original stimulus a small amount, and added a small amount of noise, prior to computing statistics. The intent of the blur was to mimic the reduction of resolution in peripheral vision (Anstis, 1974; van Essen & Anderson, 1990). The added noise allows P–S to work better on typical crowding displays with large blank background regions—such blank regions can cause instability in the algorithm, which was designed to work on more dense, texture-like images.

Each mongrel was then made by synthesizing a "texture" from the statistics computed from a given stimulus. Every stimulus gives rise to many synthesized

Crowding stimuli					
Targets	E, T, V, X	F, E, L, T	F, E, L, T	F, E, L, T	F, E, L, T
Flankers	C, G, O, Q	Curves	Lines	Letters ≠ FELT	Squiggles
Sorting stimuli (mongrels)					
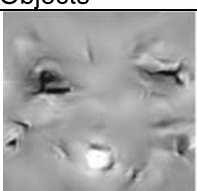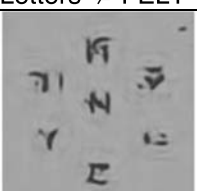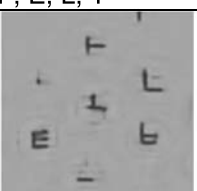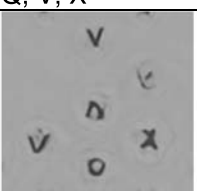
Crowding stimuli				
Targets	F, E, L, T	F, E, L, T	F, E, L, T	A, C, K, O
Flankers	Objects	Letters ≠ FELT	F, E, L, T	Q, V, X
Sorting stimuli (mongrels)				

Figure 6. Sample stimuli for the crowding and sorting experiments. The top row shows a sample stimulus from the crowding experiment. The middle two rows specify the 4 possible targets for each condition and the flankers. The bottom row shows a sample mongrel for the corresponding sorting task. Conditions are shown in order of mean difficulty on the crowding task, from easiest (top left) to most difficult (bottom right).

patterns; one can find other patches that satisfy the given constraints simply by initializing the synthesis process with a different noisy image. Typical crowding stimuli are localized in space, and the subjective percept of crowded arrays maintains some of that localization (Freeman & Pelli, 2007). To achieve this effect with Portilla and Simoncelli's algorithm, we initialize the texture synthesis process with a noisy, very blurry version of the original image (images were blurred with a Gaussian with a standard deviation of approximately half the width of target letters, see Figure 7). This initialization tends to lead to a synthesized patch with some of the very low spatial frequency characteristics of the original stimuli— the global shape and position of the array are roughly preserved in the mongrel. This procedure helps discourage the P–S algorithm from producing mongrels with individual items that "wrap around" the edges of the image (a quirk of the algorithm that is not problematic for texture applications but poses obvious difficulties here). The choice of a seed image can influence the resulting mongrels. That said, the seed image is not a constraint in the same sense as the actual measurements made by the P–S analysis procedure since the algorithm never adjusts the mongrel to ensure that any properties of the seed image are being preserved. To make clear the difference between running the algorithm with and without the initialization procedure described here, Figure 7 contains mongrels made by our procedure and by a white-noise image. In terms of difficulty identifying the target letter, the differences between the two are subtle. Figure 6 shows one sample mongrel per condition; supplementary information for this paper contains the complete set of mongrels.

Each sorting task was a direct analog of the crowding task from which stimuli were drawn to create mongrels. Participants were given a set of mongrels printed on 3″
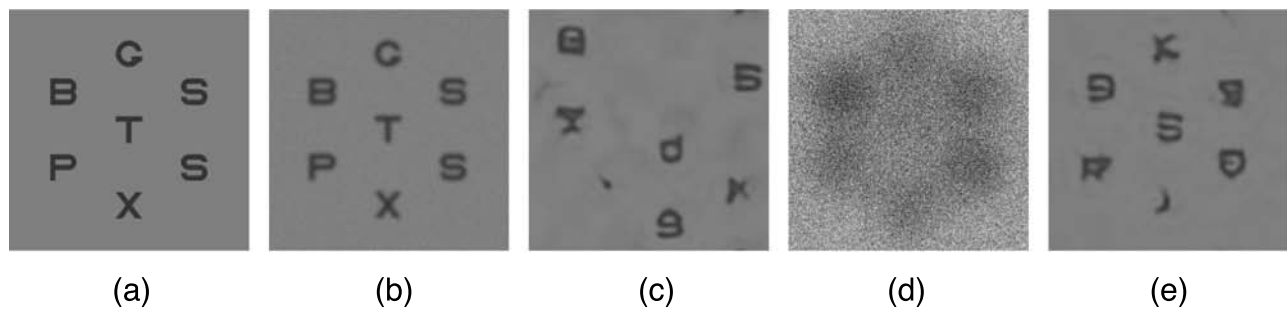
Figure 7. The mongrelization process. (a) The original crowded display. (b) Blurred original (to account for reduced acuity), with added noise; the input to the texture analysis routine. (c) Mongrel created starting with a random noise seed. (d) Seed consisting of a low-pass version of the original plus random noise, shown at 8× contrast. (e) The mongrel resulting from the seed in (d).

square cards. For each sorting task, participants could view the cards for unlimited time. Participants were told the 4 targets for each condition and shown examples of original stimuli. They were told that each mongrel resulted from "jumbling" one of the original stimuli; as a result the target would not necessarily appear in the center of the mongrel, nor look exactly like the target letter. Participants were asked to sort the cards for each condition into 4 piles, corresponding to which target they thought was contained in the corresponding original stimulus.

Pattern classification

To objectively assess the relative difficulty of our crowding tasks under our hypothesized representation, we analyzed class separability for each set of "crowded" stimuli given the summary statistics measured by the model. This is akin to replacing our human observer with a computer observer who only has access to the list of statistics measured in each original stimulus image.

The full set of images for each crowding task leads to a set of corresponding vectors of summary statistics (1384 coefficients/vector). We normalized the standard deviation of each coefficient over our entire corpus of stimuli and added a small amount of zero-mean Gaussian "observation noise" to each measurement. This single noise parameter was adjusted to best fit the data. We then used Principal Components Analysis (PCA) to embed these feature vectors in a 2-dimensional subspace for classification. Given these 2-D points, we computed the performance of a linear discriminant analysis that assigned each stimulus to a target category based on its position relative to the training data (using a leave-one-out procedure), given a circularly symmetric estimate of class variance. The classes used as labels in our discriminant analysis represent the 4 target letters that appeared in the center of each target array, making this analysis directly applicable to the 4AFC judgment required in both the crowding and mongrel-sorting tasks.

The machine learning methods used in this analysis are very simple, and we do not suggest that this analysis represents an attempt to exhaustively examine the optimal parameters for carrying out pattern classification of the P–S features. Inclusion of more dimensions in the analysis, less noisy measurements, or use of more sophisticated methods would likely yield greater absolute classification accuracy. However, the critical test of our model is not whether we can get absolute performance rates as high as possible, but whether we can successfully capture the relative difficulty of the 9 crowding tasks using only the ensemble representation of image structure. As mentioned above, with our limited amount of data it is impossible to know what kind of classifier to use. Our goal here is simply to demonstrate that machine classification will one day be feasible. To that end, we have opted here to keep our machine learning methods simple and project to a very low-dimensional space with the number of dimensions chosen so as to get approximately the same average performance in the machine classifier as we saw in the human performance of crowding tasks.

Results

First we compare accuracy at mongrel sorting to accuracy in the analogous crowding tasks. If mongrels are indeed an effective visualization of the information available in the crowded displays, we would expect difficulty in each mongrel task to be the same as the difficulty in the corresponding crowding task. Figure 8a shows a scatter-plot of sorting performance vs. crowding performance along with a best-fit regression line and a 45° line (dashed) indicating what perfect prediction would look like.

The crowding tasks we selected spanned a reasonably wide range of difficulty. Considering crowding performance relative to sorting performance, we found a significant positive correlation between tasks (Pearson's R^2 = 0.65, $p < 0.01$, one-tailed), indicating that the statistics visualized by our mongrels constrain task performance in a similar manner as crowding. By comparison, average R^2
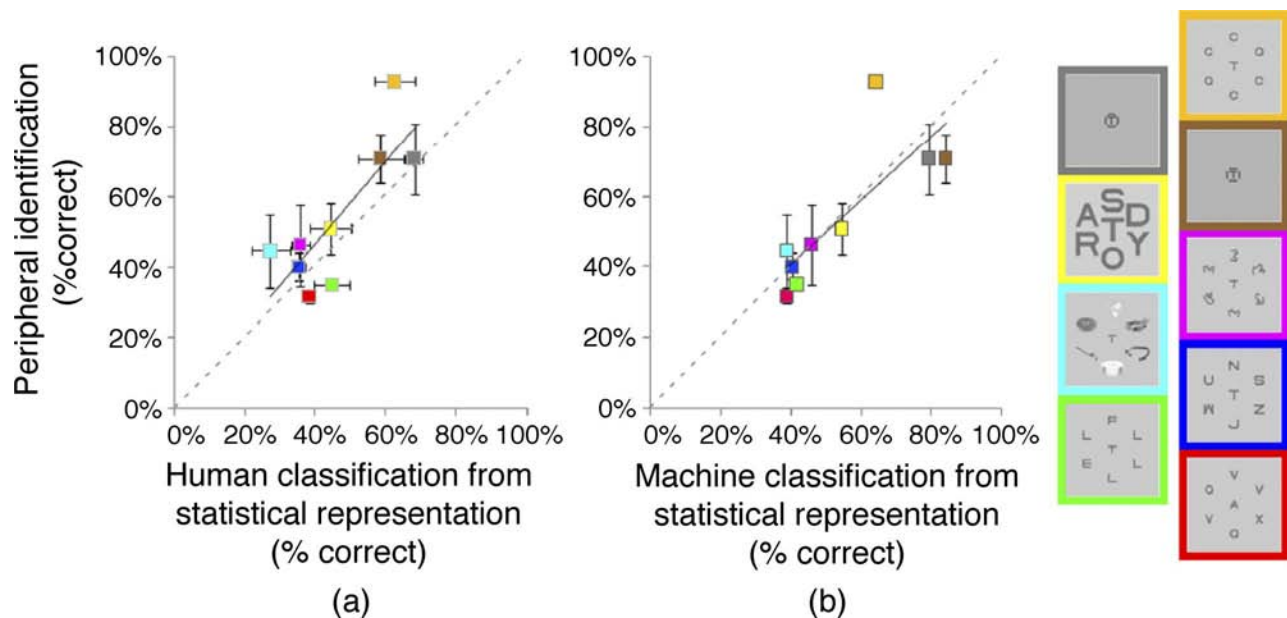
Figure 8. Each square represents a different crowding condition. The conditions, in order of difficulty, are shown in Figure 5. (a) Correlation between human performance at identifying peripheral targets under conditions of crowding tasks, and human performance sorting foveally viewed mongrels according to likely target identity. (b) Correlation between human performance identifying peripheral, crowded targets, and machine classification performance upon a vector of statistics measured in each crowded stimulus.

between subjects on the crowding task is 0.74. Furthermore, the slope of our regression line (1.2) indicates that the data is not merely correlated—this slope is not significantly different from 1 ($t(7) = 0.57$, $p > 0.20$). We take this as a key piece of evidence that mongrels capture much of the information maintained and lost under conditions of crowding.

As described above, we also compared the performance of linear classification applied to the vectors of summary statistics extracted from the crowded stimuli to human crowding performance. If our representation does in fact capture the information available under crowding, then we would expect a significant positive correlation between classifier performance and crowding performance. In Figure 8b, we display a scatter-plot of classifier accuracy vs. crowding performance along with a best-fit regression line. Similar to our findings with human observers sorting mongrelized stimuli, we observe a significant positive correlation between our machine classification results and the performance achieved by participants in a visual crowding task (Pearson's $R^2 = 0.64$, $p < 0.01$, one-tailed). Again, the slope of our regression line (0.9) is not significantly different from 1 ($t(7) = 0.34$, $p > 0.20$), indicating that machine classification is not merely well correlated with crowding performance but actually predicts it reasonably well. Though we carried out our classification procedure using a 2-D representation of the stimuli, the obtained R values obtained from embeddings with higher dimensionality do not differ greatly. For dimensionality between 3 and 10, we observed a minimum R^2 of 0.61 and a maximum of 0.66. This provides further support for our proposal that the proposed representation of crowded stimuli captures the information available to the visual system under conditions of crowding.

Discussion

Mongrels enable a useful methodology for testing perceptual models

A fundamental problem for understanding representation in the visual system lies in understanding what stimuli elicit the same response from a given model. If a "cell" (either physiological or computational) responds to a particular stimulus, e.g., a given letter, "A", what else should we expect it to respond to? This concept of equivalence classes underlies much of the study of vision. In color vision, for instance, two colors from the same equivalence class elicit the same stimulation throughout the visual system. Thus, they can be "silently" substituted for each other without an observer noticing the transition (Donner & Rushton, 1959). "Texture metamers," on the other hand, hypothetically produce the same activation in a set of "texture filters," but different activations in the retina and early parts of the visual system. As a result, texture metamers preattentively appear identical but may be discriminable with attention or when observers are allowed to view the transition between the metamers (Chubb, Nam, Bindman, & Sperling, 2007; Richards, 1979). Recent advances in texture synthesis (Heeger &

Bergen, 1995; Portilla & Simoncelli, 2000) allow the creation of texture metamers for a broader class of textures (Balas, 2006).

Work on texture metamers has used synthesized textures to study texture perception, per se. Researchers essentially compare the perception of an original texture to a synthesized one and ask whether they appear to be the same type of texture. In this paper, we ask whether a texture representation underlies vision more generally, not just tasks that explicitly involve texture perception. We hypothesize that early stages of vision extract summary statistics from the visual input, with a pooling region that grows with eccentricity. The mongrels of a given stimulus share approximately the same summary statistics as the original stimulus, constituting an equivalence class of the hypothesized model.

A lossy representation in early vision by summary statistics has profound implications for performance of visual tasks in general. Mongrels enable the study of this representation while respecting the richness of the visual tasks likely to be affected—the symbol identification tasks common in the visual crowding literature constitute just one example. By providing equivalence classes of our model—essentially allowing visualization of the information available in a given set of summary statistics—mongrels allow us to sidestep the need for a model of human object/pattern recognition. Human observers view mongrels and perform a given task. This methodology allows us to test whether the information loss inherent in a low-level representation via summary statistics predicts task performance, without needing to understand higher level visual processing. This technique is not limited to the model we have proposed; any model that can be represented by a transformation of the visual input, or by a set of constraints on the visual input, can be tested in this fashion. Phenomena to which this technique applies extend beyond visual crowding.

As an added example of the utility of visualizing the information available in the hypothesized representation, consider the array shown in Figure 9a, consisting of four A's. Unlike the ABAB array of Figure 2, the AAAA array, viewed peripherally, seems to contain letter-like

forms that, while jumbled, are unmistakably A's. Figure 9b shows a sample mongrel for this stimulus. The forms in this mongrel all look A-like, in accord with subjective impressions of Figure 9a in the periphery.

Note, however, that the AAAA mongrel contains what looks like an inverted A. There is only a small difference between the statistics (e.g., orientations and their relative arrangement) of an array of pure A's and an array also containing inverted A's. Is this a "bug" resulting from faults in the underlying model, or does it accurately reflect the nature of the peripheral percept?

In a pilot experiment, we asked 8 subjects to view crowded arrays of A's and report if an inverted A was present in each array. All stimuli were 2×2 arrays of A's, with asterisks on the border of the entire array, to prevent attending to the less crowded edges of the array to do the task. All "upright" stimuli contained 4 upright A's. "Inverted" stimuli contained a single inverted A, randomly assigned to one of the 4 positions in the array. Subjects viewed 16 trials in each condition, fixating on a centrally presented cross while stimulus arrays were presented at approximately 10 degrees eccentricity for 250 ms. Individual A's and asterisks subtended approximately 1.5 degrees, and center-to-center spacing between elements was approximately 1 degree. Trial order was randomized and no feedback was provided.

Observers' performance at discriminating between the "all upright" and "inverted" conditions was not significantly different from chance (mean accuracy 55%, $N = 8$, $p > 0.05$). Introspection on the mongrels corresponding to a given stimulus can thus suggest interesting testable predictions about peripheral perception.

Representation by summary statistics effectively predicts performance in the periphery

We have presented a concrete proposal that peripheral stimuli are represented in the visual system by local summary statistics and suggested a candidate set of statistics that might be computed. We have shown,

(a)   (b)
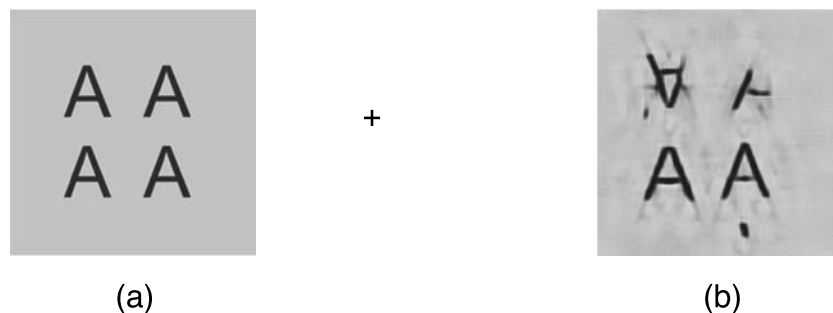
Figure 9. The letters in (a) are within the critical spacing for crowding, when fixating on the "+". (b) A sample mongrel. Note that (b) contains an inverted "A" despite a homogeneous input of upright A's. This unexpected result predicts that it should be difficult to distinguish homogenous arrays from those containing inverted A's under crowding, which was confirmed by pilot data.

through a mongrel-sorting task, that performance with the summary statistics predicts performance at peripheral crowding tasks, suggesting that in fact we have captured the information available in the periphery. Of course, further refining our understanding of visual representation will require an even greater diversity of stimuli and tasks. In doing this, we will be greatly aided by the fact that we can generate a mongrel of a patch of any arbitrary stimulus.

Ultimately, we would like to get the human out of the loop; we would like to be able to give the model labeled sample images for a given task and have the model predict performance at that task. Supervised learning for the particular task would allow the model to predict performance based upon the measured statistical vectors. At present, this is difficult, as we have insufficient data to replace the "normal observer" in Figure 4 with a machine classifier. However, we have shown preliminary results indicating the plausibility of this approach. By making

reasonable choices for a machine classifier, we were able to use it to get reasonable predictions of crowding performance.

Our model also provides for continuity between crowded perception and uncrowded perception. The fundamental puzzle of crowding is that observers can easily identify simple stimuli such as letters in the periphery when they are presented in isolation, yet may be unable to identify those same stimuli when flankers appear within the critical spacing. A good model of crowding should also predict this *lack* of crowding. Figures 10b and 10c show sample mongrels for an isolated letter stimulus (Figure 10a). Note that in the mongrels for Figure 10a, the letters are clearly identifiable, and that the equivalence class of the model for these stimuli consists of similar letters at somewhat different spatial locations. This result suggests that there is a unified representation for peripheral vision—our model does not require some
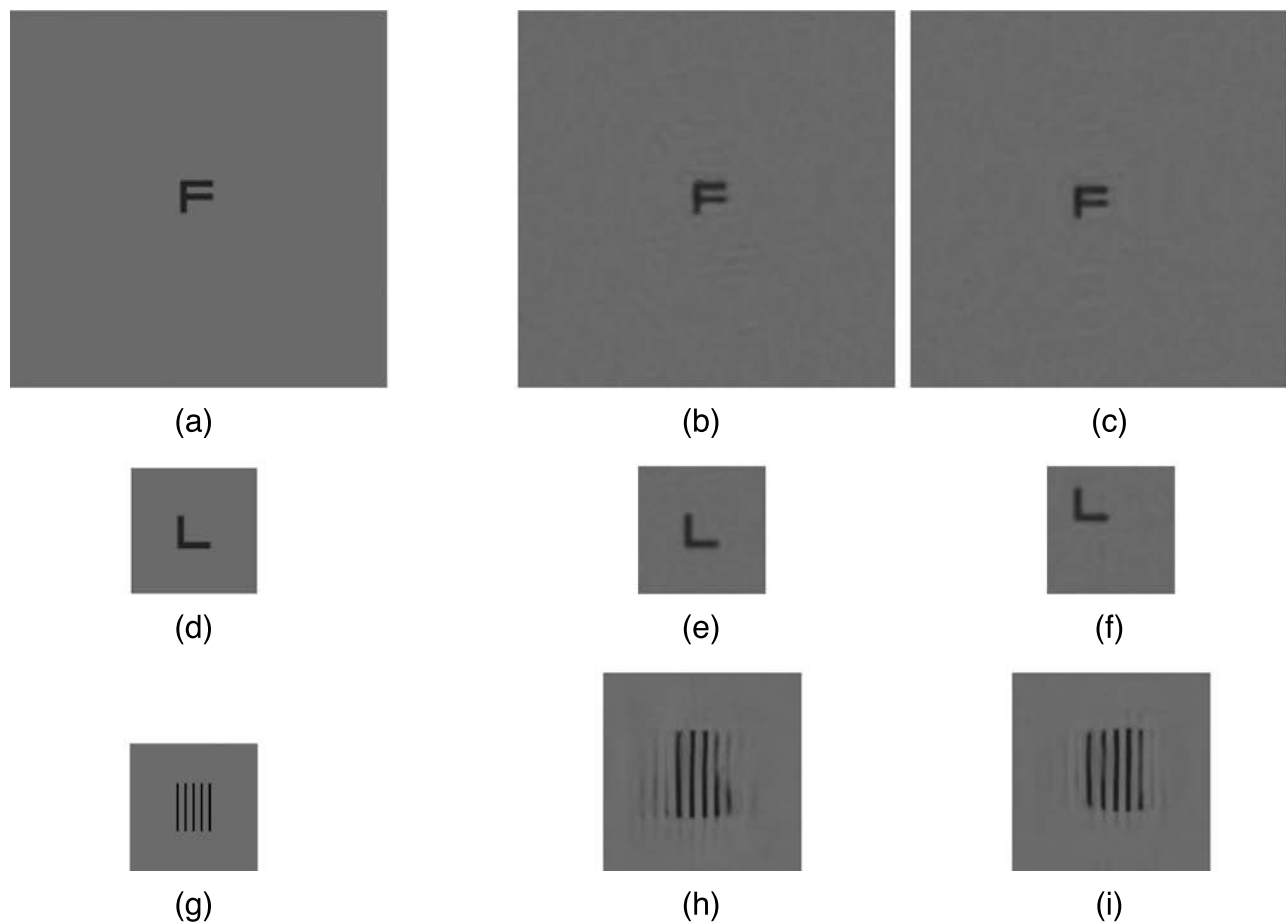

Figure 10. (a) An uncrowded "F", viewed peripherally (large pooling region). (b, c) Associated mongrels, sharing the same joint statistics as the original (a), and thus belonging to the equivalence class of (a), according to our model. (d) A simple stimulus, with a small pooling region (the size of the image), e.g., as in foveal viewing. Members of the equivalence class are shown by the mongrels in (e) and (f). Note that in both cases (a, d), the statistics are sufficient to describe the stimulus up to a translation. Thus the model can likely identify isolated stimulus letters in the periphery, as well as in the "fovea." (g) Original image of 5 bars, and (h, i) associated mongrels. Note the apparent difference in the number of bars in (h) and (i), and the ghostly extra bars flanking the higher contrast ones. The mongrels correctly predict uncertainty about the number of bars present. Thus our model can even predict where "normal" object recognition breaks in the fovea. (The mongrels in (h) and (i) have been enlarged for clearer viewing.)

component that "turns on" in the presence of flankers and makes vision difficult.

Stimulus complexity relative to pooling region size determines the effectiveness of summary statistics in representing the stimulus

For more complex objects, even without flankers summary statistics may not provide enough information to perform a given identification task. Recent work in computer vision has in fact suggested that certain object recognition tasks can be performed using texture-like descriptors, often known as a "bag of words" or "bag of features," while for other tasks this descriptor is insufficient (Lowe, 2004). The key to whether a particular task is easy or difficult is essentially the complexity of the stimuli within the pooling region. If a stimulus is too complex within a pooling region, then computed statistics will provide a poorer representation of that stimulus (see van den Berg, Cornelissen, & Roerdink, 2009 for related work regarding visual clutter). This may be related to the empirical phenomenon Martelli et al. (2005) have dubbed "self-crowding." Visualizing the equivalence classes of our model should generate rich predictions about what stimuli and tasks lead to self-crowding. In fact, what does and does not self-crowd should inform particular choices of statistics computed by the model.

Bouma's law says pooling regions get smaller as you move into foveal vision, so it should be easier to faithfully represent stimuli in foveal vision, even with the same representation we propose here. A smaller pooling region may isolate a single letter from its flankers, and thus enable letter recognition. Figure 10d shows a toy example, in which an "L" has been isolated by a smaller pooling region. Figures 10e and 10f show two mongrels derived from this stimulus. Similar to the uncrowded peripheral letter in Figures 10a–10c, the L is represented by up to a small tolerance in position. Nonetheless, it is still possible within our model for a stimulus to be too complex for its pooling region, even in the fovea (Figures 10g–10j). Behaviorally, one can also get some crowding-like effects for stimuli like this one, even in the fovea (Ehlers, 1936).

Toward a general model of early visual representation: What and why?

As discussed above, when pooling regions are "too large" for the complexity of the stimuli, our model behaves as if it is collecting summary statistics; there is a loss of information coupled with degraded performance at a task. It seems promising that when the pooling regions are small relative to stimulus complexity, the model behaves more like early stages of object recognition: it can represent not only oriented structures like bars but also piece together such simple structures to represent

more complicated features such as corners and simple letters. The model gains sensitivity to more complicated structures at the same time as it acquires tolerance to small changes in position of the pattern.

This raises the more ambitious question of whether a single model could capture much of early visual representation, simply by increasing the size of the pooling region with eccentricity. Comparison with standard feed-forward models of object recognition (e.g., Fukushima, 1980; Riesenhuber & Poggio, 1999) attests to the plausibility of this hypothesis. Early stages of these models 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. These co-occurrences, computed over a larger pooling region, can approximate the *correlations* computed by our model.

A unified representation in early vision in terms of summary statistics may provide the visual system with an effective strategy for dealing with the bottleneck in visual information processing. This representation compresses the visual input, while allowing normal object recognition where pooling regions are small. Even with the loss of information over larger pooling regions, mongrels show that summary statistics can capture a great deal of information about the stimulus: its sharpness and spatial frequency content; the presence of extended structures, junctions, and curves; the homogeneity of the pattern, and so on. In fact, in our normal visual experience, we feel like we have a rich representation of the visual world, and not just at the fovea. We are puzzled when introspection or experiments reveal that we are unaware of the details. The puzzle, perhaps, comes from thinking: if we have a "picture" of the visual world—an image-based representation—why can we not access the details? However, the proposed representation is *not* image-like but rather consists of a number of measurements of local summary statistics. At a given moment, for a given fixation, many details will be lost due to the loss of information inherent in this encoding. Nonetheless, local summary statistics provide a rich description of the visual world, by allowing us to perform such tasks as image segmentation (Balas, 2006; Beck, 1983; Julesz, 1981; Keeble et al., 1995; Rosenholtz, 2000; Voorhees & Poggio, 1988), detecting unusual items (Rosenholtz, 1999, 2001), making judgments about materials (Motoyoshi et al., 2007), judging the gist of a scene (Greene & Oliva, 2009; Oliva & Torralba, 2001; Renninger & Malik, 2004), and, likely, directing eye movements.

Conclusions

We have proposed a model in which the visual system represents the visual input via local summary statistics. To test this hypothesis, a key component of our methodology is visualizing the information encoded by a given set of

statistics. Given any stimulus, there is a set of images consistent with the statistics of the original image (the equivalence class). Our hypothesis suggests that peripheral vision implicitly represents this distribution of images rather than a particular stimulus. Using texture synthesis techniques, we can generate samples from the putative distribution and show them foveally to predict the confusions that subjects will make. Our approach can be utilized with arbitrary images, and not just stimuli drawn from a specific set such as letters or gratings.

We have demonstrated that our model can predict performance on peripheral identification tasks in which the target is crowded due to flankers within the pooling region. Our results indicate the feasibility of a unified account of representation in peripheral vision (crowded and not), texture perception, and ordinary object recognition. The intriguing phenomenon of visual crowding may be a natural result of a successful strategy for dealing with a bottleneck in visual processing; representing the visual input via summary statistics allows the visual system to simultaneously reduce the information passing through the bottleneck, and yet encode a great deal of useful information about the visual world.

Appendix A

Here we describe the statistics computed by our model in more detail. As mentioned in the main body of this paper, we have here used the statistics from Portilla and Simoncelli's (2000) texture analysis/synthesis routine as our candidate representation in peripheral vision. While we are not wedded to these particular statistics, our results show that the information captured by these statistics constrains task performance in a manner similar to visual crowding. Our goal in this appendix is to provide the reader with a more complete intuition of the information

captured by these statistics. For a full description of the statistics used, as well as more examples of the information captured by these statistics, we refer the interested reader to Portilla and Simoncelli (2000)).

The full P–S algorithm involves the measurement of 1384 parameters obtained primarily from a wavelet-based pyramid decomposition of the input image. (In this paper, we use a wavelet pyramid with 4 orientations and 4 scales, and a neighborhood of size 9.) This multitude of measurements can be grouped into four distinct families of statistics: Pixel statistics, correlation coefficients, magnitude correlations, and relative phase. Below we include a brief description of each distinct family of statistics.

These statistics interact in complex ways in constraining the set of possible images, and as a result it is difficult to simply intuit what information the statistics encode based upon their description. To aid in intuitions, here we show, for each family of statistics, an example of a texture synthesized *without* that family of statistics. By looking at the information that seems to be *missing* from these "lesioned" textures, when compared to textures synthesized with the full set of statistics (Figure A1b), one can get better insights into the information encoded by a given family of statistics. We use as our example an image taken from the Brodatz database (Figure A1a; Brodatz, 1996). Balas (2006) used just this sort of lesioning of the model to test the importance of the various classes of statistics for parafoveal texture discrimination. The reader may want to get an intuition for Balas' results by comparing the original texture with each lesioned texture in peripheral vision.

Pixel statistics

The first set of statistics characterizes the pixel intensity distribution of the target image. While of course the visual system knows nothing about "pixels," these statistics allow representation of the distribution of the raw intensities present in the original stimulus. The Portilla–
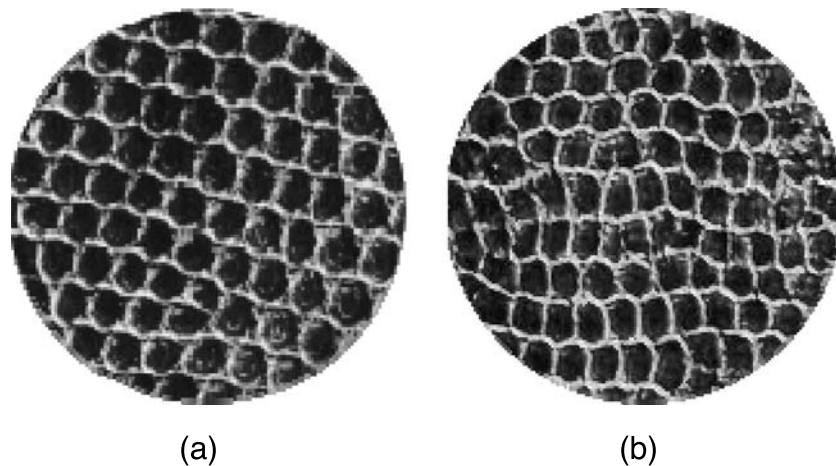

(a)   (b)

Figure A1. (a) A sample texture drawn from the Brodatz database. This is a sample of reptile skin. (b) Reptile skin synthesized using the full set of statistics from Portilla and Simoncelli (2000).

Simoncelli algorithm does not maintain the entire intensity histogram of the input image but instead maintains moments of this histogram that are sufficient to accurately capture (for the purposes of texture synthesis, at least) the correct intensities in a given image. Specifically, the intensity histogram is represented by the mean, variance, kurtosis, skew, and range (minimum and maximum values) of intensity. The skew and kurtosis of a low-pass version of the input image is also calculated.

In Figure A1b, we display a portion of synthesized reptile skin texture created with the full P–S statistics. Figure A2a shows a portion of a synthesized texture created without constraining these pixel statistics. The lack of sufficient constraint on the pixel intensities has large consequences for the appearance of the synthesized texture, particularly for the contrast.

Correlation coefficients

Periodicity, or the degree to which a pattern repeats, is an important component to the appearance of a pattern, particularly of a texture. P–S captures repeated image structure via the local autocorrelation function (ACF) of low-pass versions of the stimulus. Figure A2b displays a texture synthesized without these measurements of local periodicity. Clearly this or equivalent information is required to adequately capture the strong periodic component of this image.

Magnitude correlations

This set of statistics records the co-occurrence of responses of oriented wavelets across several different kinds of "neighbors." This set of statistics essentially records edge co-occurrence across position (does a particular edge continue in a straight line or a corner?), across orientation (do vertical edges tend to co-occur with horizontal edges?), and across scale (is a small-scale edge part of a larger structure at a coarser scale?). These statistics are a powerful means of describing a wide range of structures within the image including extended contours that are straight and curved, closed contours, and corners within the image. Figure A2c displays a texture synthesized without any constraint on these parameters, giving rise to distinctive errors in the synthetic image. Balas (2006) showed that these statistics were very important to parafoveal texture appearance.

Relative phase statistics

This last family of statistics measures the relative phase of wavelet features between neighboring spatial scales within the pyramid decomposition of the target image. This is a measure of the dominant local relative phase between coefficients within a subband and their neighbors in the immediately larger scale. Figure A2d displays a lesioned texture in which these statistics have not been constrained. Note how this gives rise to interesting reversals of polarity across the image.

We encourage interested readers to download the MATLAB implementation of the Portilla–Simoncelli algorithm and explore the nature of the synthetic images created from different kinds of stimuli. While our descriptions here provide some visual intuitions regarding the functionality of the model, hands-on experimentation with the model is extremely rewarding.

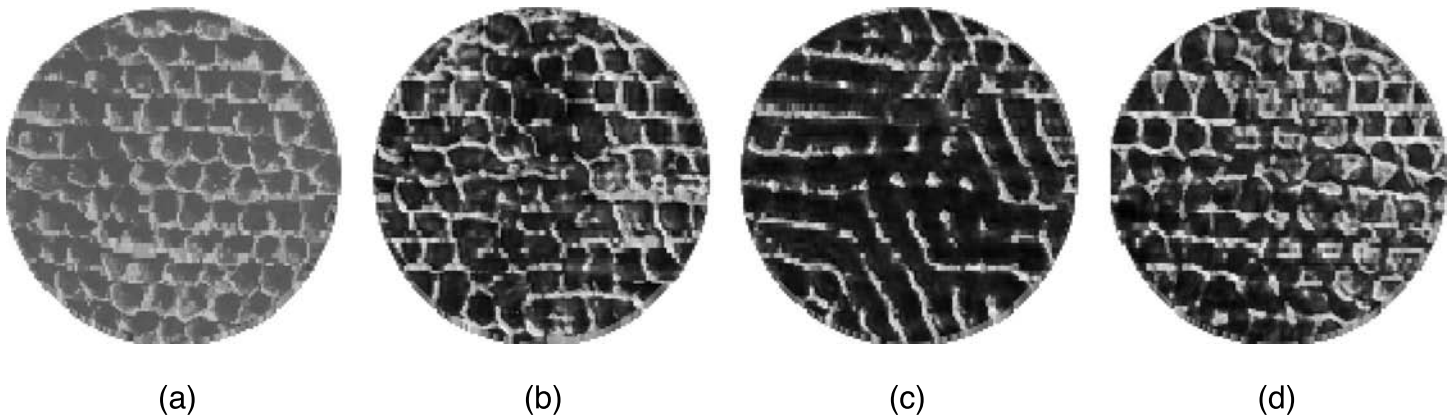
(a)   (b)   (c)   (d)

Figure A2. Synthesized reptile skin. Each of these samples was synthesized *without* constraining one of the sets of statistics from Portilla and Simoncelli (2000). Compare each sample with Figure A1b to get a sense of what information is captured by the missing statistics. (a) No constraint on the distribution of pixel intensities. The overall appearance of the texture is fairly accurate, but the contrast and brightness are obviously incorrect. (b) No constraint on the local periodicity (autocorrelation). While much of the structure is still evident (individual cells are still observable, for example) the global arrangement of cells into a repeated pattern is less evident. (c) No constraint on the magnitude correlation statistics. Here, while some aspects of the periodic structure are evident, the structure of individual cells is not preserved. (d) No constraint on the relative phase statistics. Note the contrast-polarity errors throughout the image: some cells have a light interior while others have a black interior.

Acknowledgments

Footnotes

[1]Strictly speaking, a "statistic" can be any function of the data. However, in practice statistics summarize the data in some way: mean, variance, 95th percentiles, full first-order histograms, full second-order histograms, correlations, and so on. It is this typical usage of "statistic" to which we refer throughout this paper.

[2]The visualization of marginal statistics in Figure 5c uses a version of Heeger–Bergen texture synthesis, in which coarser scales are not subsampled to create a pyramid. This oversampled Heeger–Bergen produces better syntheses of these stimuli than standard Heeger–Bergen and represents our best attempts to synthesize Figure 5a using only marginal statistics.

References

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research, 16,* 71–78. [PubMed] [Article]

Anstis, S. M. (1974). Letter: A chart demonstrating variations in acuity with retinal position. *Vision Research, 14,* 589–592. [PubMed]

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

Beck, J. (1983). Textural segmentation, second-order statistics, and textural elements. *Biological Cybernetics, 48,* 125–130. [PubMed]

Brettel, H., Viénot, F., & Mollon, J. D. (1997). Computerized simulation of color appearance for dichromats. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 14,* 2647–2655. [PubMed]

Brodatz, P. (1996). *Textures: A photographic album for artists and designers.* New York: Dover.

Chubb, C., Nam, J.-H., Bindman, D. R., & Sperling, G. (2007). The three dimensions of human visual sensitivity to first-order contrast statistics. *Vision Research, 47,* 2237–2248. [PubMed] [Article]

Chung, S. T., Levi, D. M., & Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research, 41,* 1833–1850. [PubMed] [Article]

Donk, M. (1999). Illusory conjunctions are an illusion: The effects of target–nontarget similarity on conjunction and feature errors. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1207–1233.

Donner, K. O., & Rushton, W. A. (1959). Retinal stimulation by light substitution. *The Journal of Physiology, 149,* 288–302. [PubMed] [Article]

Ehlers, H. (1936). The movements of the eyes during reading. *Acta Ophthalmologica, 14,* 56.

Eriksen, C. W., & Hoffman, J. E. (1973). The extent of processing of noise elements during selective encoding from visual displays. *Perception & Psychophysics, 14,* 155–160.

Estes, W. K. (1982). Similarity-related channel interactions in visual processing. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 353–382. [PubMed]

Flom, M. C. (1991). Contour interaction and the crowding effect. *Problems in Optometry, 3,* 237–257.

Freeman, J., & Pelli, D. G. (2007). An escape from crowding. *Journal of Vision, 7*(2):22, 1–14, http://journalofvision.org/7/2/22/, doi:10.1167/7.2.22. [PubMed] [Article]

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202. [PubMed]

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58,* 137–176. [PubMed] [Article]

Heeger, D., & Bergen, J. (1995). Pyramid-based texture analysis/synthesis. *Proceedings of the 22nd Annual Conference on Computer Graphics & Interactive Techniques, 30,* 229–238.

Ivry, R. B., & Prinzmetal, W. (1991). Effect of feature similarity on illusory conjunctions. *Perception & Psychophysics, 49,* 105–116. [PubMed]

Julesz, B. (1981). A theory of preattentive texture discrimination based on first-order statistics of textons. *Biological Cybernetics, 41,* 131–138. [PubMed]

Keeble, D. R. T., Kingdom, F. A. A., Moulden, B., & Morgan, M. J. (1995). Detection of orientationally multimodal textures. *Vision Research, 35,* 1991–2005. [PubMed] [Article]

Kooi, F. L., Toet, A., Tripathy, S. P., & Levi, D. M. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial Vision, 8,* 255–279. [PubMed]

Levi, D. M., Hariharan, S., & Klein, S. A. (2002). Suppressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *Journal of Vision, 2*(2):3, 167–177, http://journalofvision.org/2/2/3/, doi:10.1167/2.2.3. [PubMed] [Article]

Loomis, J. M. (1978). Lateral masking in foveal and eccentric vision. *Vision Research, 18,* 335–338. [PubMed]

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60,* 91–110.

Martelli, M., Majaj, N. J., & Pelli, D. G. (2005). Are face processed like words? A diagnostic test for recognition by parts. *Journal of Vision, 5*(1):6, 58–70, http://journalofvision.org/5/1/6/, doi:10.1167/5.1.6. [PubMed] [Article]

"mongrel." Merriam-Webster Online Dictionary. (2008). Merriam-Webster Online. 12 June 2008, http://www.merriam-webster.com/dictionary/mongrel.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447,* 206–209. [PubMed]

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

Nazir, T. A. (1992). Effects of lateral masking and spatial precueing on gap-resolution in central and peripheral vision. *Vision Research, 32,* 771–777. [PubMed] [Article]

Oliva, A., & Torralba, A. (2001). Modeling the shape of a scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175.

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

Peli, E., & Geri, G. (1999). Testing the simulation of peripheral vision with image discrimination. *Technical Digest of Papers, SID-99, Digest,* 424–427.

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://journalofvision.org/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,* 1129–1135. [PubMed]

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40,* 49–71.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44,* 2301–2311. [PubMed] [Article]

Richards, W. (1979). Quantifying sensory channels: Generalizing colorimetry to orientation and texture, touch, and tones. *Sensory Processes, 3,* 207–229. [PubMed]

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025. [PubMed]

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

Rosenholtz, R. (2000). Significantly different textures: A computational model of pre-attentive texture segmentation. In Vernon (Ed.), *Proceedings of the European Conference on Computer Vision, LNCS 1843* (pp. 197–211). Dublin, Ireland: Springer Verlag.

Rosenholtz, R. (2001). Search asymmetries? What search asymmetries? *Perception & Psychophysics, 63,* 476–489. [PubMed] [Article]

Strasburger, H., Harvey, L. O., Jr., & Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Perception & Psychophysics, 49,* 495–508. [PubMed]

van den Berg, R., Cornelissen, F. W., & Roerdink, J. B. (2009). A crowding model of visual clutter, *Journal of Vision, 9*(4):24, 1–11, http://journalofvision.org/9/4/24/, doi:10.1167/9.4.24. [PubMed] [Article]

van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision, 7*(2):14, 1–11, http://journalofvision.org/7/2/14/, doi:10.1167/7.2.14. [PubMed] [Article]

van Essen, D. C., & Anderson, C. H. (1990). Information processing strategies and pathways in the primate retina and visual cortex. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *Introduction to neural & electronic networks* (pp. 43–72). San Diego, CA: Academic Press Professionals, Inc.

Virsu, V., & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification

factor. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, 37,* 475–494. [PubMed]

Voorhees, H., & Poggio, T. (1988). Computing texture boundaries from images. *Nature, 333,* 364–367. [PubMed]

Wilkinson, F., Wilson, H. R., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 14,* 2057–2068. [PubMed]