Texture perception
Ruth Rosenholtz
Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, CSAIL, Cambridge, MA, USA

To appear in:
Oxford Handbook of Perceptual Organization
Oxford University Press
Edited by Johan Wagemans

Abstract
Texture informs our interpretation of the visual world. It provides a cue to the shape and orientation of a surface, to segmenting an image into meaningful regions, and to classifying those regions, e.g. in terms of material properties. This chapter discusses recent advances in understanding of segmentation and representation of visual texture. Successful models have described texture by a rich set of image statistics (“stuff”) rather than by the features of discrete, pre-segmented texture elements (“things”). Texture processing mechanisms may also underlie important phenomena in peripheral vision known as crowding. If true, such mechanisms would influence the information available for object recognition, scene perception, and many visual-cognitive tasks.

Keywords: texture, texture segmentation, representation, crowding, image statistics, stuff vs. things

1. Introduction: What is texture?

The structure of a surface, say of a rock, leads to a pattern of bumps and dips which we can feel with our fingers. This applies equally well to the surface of skin, the paint on the wall, the surface of a carrot, or the bark of a tree. Similarly, the pattern of blades of grass in a lawn, pebbles on the ground, or fibers in woven material, all lead to a tactile “texture”. The surface variations that lead to texture we can feel also tend to lead to variations in the intensity of light reaching our eyes, producing what is known as “visual texture” (or here, simply “texture”). Visual texture can also come from variations that do not lend themselves to tactile texture, such as the variation in composition of a rock (quartz looks different from mica), waves in water, or patterns of surface color such as paint.

Texture is useful for a variety of tasks. It provides a cue to the shape and orientation of a surface (Gibson 1950). It aids in identifying the material of which an object or surface is made (Gibson 1986). Most obviously relevant for this Handbook, texture similarity provides one cue to perceiving coherent groups and regions in an image.

Understanding human texture processing requires the ability to synthesize textures with desired properties. By and large this was intractable before the wide availability of computers. Gibson (1950) studied shape-from-texture by photographing wallpaper from different angles. Our understanding of
texture perception would be quite limited if we were restricted to the small set of textures found in wallpaper. Attneave (1954) gained significant insight into visual representation by thinking about perception of a random noise texture, though he had to generate that texture by hand, filling in each cell according to a table of random numbers. Beck (1966; 1967) formed micropattern textures out of black tape affixed to white cardboard, restricting the micropatterns to those made of line segments. Olson and Attneave (1970) had more flexibility, as their micropatterns were drawn in india ink. Julesz (1962, 1965) was in the enviable position of having access to computers and algorithms for generating random textures. More recently, texture synthesis techniques have gotten far more powerful, allowing us to gain new insights into human vision.

It is elucidating to ask why we label the surface variations of tree bark “texture,” and the surface variations of the eyes, nose, and mouth “parts” of a face object, or objects in their own right. One reason for the distinction may be that textures have different identity-preserving transformations than objects. Shifting around regions within a texture does not fundamentally change most textures, whereas swapping the nose and mouth on a face turns it into a new object (see also Behrmann et al., this volume). Two pieces of the same tree bark will not look exactly the same, but will seem to be the same “stuff”, and therefore swapping regions has minimal effect on our perception of the texture. Textures are relatively homogeneous, in a statistical sense, or at least slowly varying. Fundamentally, texture is statistical in nature, and one could argue that texture is stuff that is more compactly represented by its statistics – its aggregate properties – than by the configuration of its parts (Rosenholtz 1999).

That texture and objects have different identity-preserving transformations suggests that one might want to perform different processing on objects than on texture. In the late 1990s, that was certainly the case in computer vision and image processing. Object recognition algorithms differed greatly from texture classification algorithms. Algorithms for determining object shape and pose were very different from those that found the shape of textured surfaces. In image coding, regions containing texture might be compressed differently than those dominated by objects (Popat and Picard 1993). The notion of different processing for textures vs. objects was prevalent enough that several researchers developed algorithms to find regions of texture in an image, though this was hardly a popular idea (Karu et al. 1996; Rosenholtz 1999).

However, exciting recent work (Section 4) suggests that human vision employs texture processing mechanisms even when performing object recognition tasks in image regions not containing obvious “texture”. The phenomena of visual crowding provided the initial evidence for this hypothesis. However, if true, such mechanisms would influence the information available for object recognition, scene perception, and diverse tasks in visual cognition.

This chapter reviews texture segmentation, texture classification/appearance, and visual crowding. It is obviously impossible to fully cover such a diversity of topics in a short chapter. The material covered will focus on computational issues, on the representation of texture by the visual system, and on connections between the different topics.
2. Texture segmentation

2.1 Phenomena
An important facet of vision is the ability to perform “perceptual organization,” in which the visual system quickly and seemingly effortlessly transforms individual feature estimates into perception of coherent regions, structures, and objects. One cue to perceptual organization is texture similarity. The visual system uses this cue in addition to and in conjunction with (Giora and Casco 2007; Machilsen and Wagemans 2011) grouping by proximity, feature similarity, and good continuation (see also Brooks, this volume; Elder, this volume).

The dual of grouping by similar texture is important in its own right, and has, in fact, received more attention (see also Dakin, this volume). In “preattentive” or “effortless” texture segmentation two texture regions quickly and easily segregate – in less than 200 milliseconds. Observers may perceive a boundary between the two. Figure 1 shows several examples. Like contour integration and perception of illusory contours, texture segmentation is a classic Gestalt phenomenon. The whole is different than the sum of its parts (see also Wagemans, this volume), and we perceive region boundaries which are not literally present in the image (Figure 1abc).

![Figure 1. Texture segmentation pairs. (a)-(d): Micropattern textures. (a) Easily segments, and the two textures have different 2nd order pixel statistics; (b) Also segments fairly easily, yet the textures have the same 2nd order statistics; (c) Different 2nd-order](image-url)
statistics, does not easily segment, yet it is easy to tell apart the two textures; (d) Neither segments nor is it easy to tell apart the textures. (e,f) Pairs of natural textures. The pair in (f) is easier to segment, but all 4 textures are clearly different in appearance.

Researchers have taken performance under rapid presentation, often followed by a mask, as meaning that texture segmentation is preattentive and occurs in early vision (Julesz 1981; Treisman 1985). However, the evidence for both claims is somewhat questionable. We do not really understand in what way rapid presentation limits visual processing. Can higher-level processing not continue once the stimulus is removed? Does fast presentation mean preattentive? (See also Gillebert & Humphreys, this volume.) Empirical results have given conflicting answers. Mack et al. (1992) showed that texture segmentation was impaired under conditions of inattention due to the unexpected appearance of a segmentation display during another task. However, the segmentation boundaries in their stimuli aligned almost completely with the stimulus for the main task: two lines making up a large “+” sign. This may have made the segmentation task more difficult. Perhaps judging whether a texture edge occurs at the same location as an actual line requires attention. Mack et al. (1992) demonstrated good performance at texture segmentation in a dual-task paradigm. Others (Braun and Sagi 1991; Ben-Av and Sagi 1995) show similar results for a singleton-detection task they refer to as texture segregation. Certainly performance with rapid presentation would seem to preclude mechanisms which require serial processing of the individual micropatterns which make up textures like those in Figure 1a-e.

Some pairs of textures segment easily (Figure 1ab), others with more difficulty (Figure 1c). Some texture pairs are obviously different, even if they do not lead to a clearly perceived segmentation boundary (Figure 1d), whereas other texture pairs require a great deal of inspection to tell the difference (Figure 1e). Predicting the difficulty of segmenting any given pair of textures provides an important benchmark for understanding texture segmentation. Researchers have hoped that such understanding would provide insight more generally into early vision mechanisms, such as what features are available preattentively.

2.2 Statistics of pixels
When two textures differ sufficiently in their mean luminance, segmentation occurs (Boring 1945; Julesz 1962). The same seems true for other differences in the luminance histogram (Julesz 1962; Julesz 1965; Chubb et al. 2007). In other words, a sufficiently large difference between two textures in their 1st-order luminance statistics leads to effortless segmentation. Differences in 1st-order chrominance statistics also support segmentation (e.g. Julesz 1965).

However, differences in 1st-order pixel statistics are not necessary for texture segmentation to occur. Differences in line orientation between two textures are as effective as differences in brightness (Beck 1966; Beck 1967; Olson and Attneave 1970). Consider micropattern textures formed of line segments

---

1 Terminology in the field of texture perception stands in a confused state. “1st- and 2nd-order” can refer to (a) 1st-order histograms of features vs. 2nd-order correlations of those features; (b) statistics involving a measurement to the first power (e.g. the mean) vs. a measurement to the power of 2 (e.g. the variance) – i.e. the 1st- and 2nd-moments from mathematics; or (c) a model with only one filtering stage, vs. a model with a filtering stage, a non-linearity, and then a 2nd filtering stage. This chapter uses the first definition.
(e.g. Figures 1a-c). Differences in the orientations of the line segments predict segmentation better than either the orientation of the micropatterns, or their rated similarity. An array of upright Ts segments poorly from an array rotated by 90 degrees; the line orientations are the same in the two patterns. A T appears more similar to a tilted (45˚) T than to an L, but Ts segment from tilted-Ts more readily than they do from Ls.

Julesz (1965) generated textures defined by Markov processes, in which each pixel depends probabilistically on its predecessors. He observed that one could often see within these textures clusters of similar brightness values. For example, such clusters might form horizontal stripes, or dark triangles. Julesz suggested that early perceptual grouping mechanisms might extract these clusters, and that, “As long as the brightness value, the spatial extent, the orientation and the density of clusters are kept similar in two patterns, they will be perceived as one”.

It is tempting to observe clusters in Julesz’ examples and conclude that extraction of “texture elements” (a.k.a. texels), underlies texture perception. However, texture perception might also be mediated by measurement of image statistics, with no intermediate step of identifying clusters. The stripes and clusters in Julesz’ examples were, after all, produced by random processes. As Julesz (1975) put it, “[10 years ago,] I was skeptical of statistical considerations in texture discrimination because I did not see how clusters of similar adjacent dots, which are basic for texture perception, could be controlled and analyzed by known statistical methods… In the intervening decade much work went into finding statistical methods that would influence cluster formation in desirable ways. The investigation led to some mathematical insights and to the generation of some interesting textures.”

The key, for Julesz, was to figure out how to generate textures with desired clusters of dark and light dots, while controlling their image statistics. With the help of collaborators Gilbert, Shepp, and Frisch (acknowledged in Julesz 1975), Julesz proposed simple algorithms for generating pairs of micropattern textures with the same 1st- and 2nd-order pixel statistics. For Julesz’ black and white textures, 1st-order statistics reduce to the fraction of black dots making up the texture. 2nd-order or dipole statistics can be measured by dropping “needles” onto a texture, and observing the frequency with which both ends of the needle land on a black dot, as a function of needle length and orientation. Such 2nd-order statistics are equivalent to the power spectrum.

Examination of texture pairs sharing 1st- and 2nd-order pixel statistics led to the now-famous “Julesz conjecture”: “Whereas textures that differ in their first- and second-order statistics can be discriminated from each other, those that differ in their third- or higher-order statistics usually cannot” (Julesz 1975). This theory predicted a number of results, for both random noise and micropattern-based textures. For instance, the textures in Figure 1a differ in their 2nd-order statistics, and readily segment, whereas the textures in Figure 1d share 2nd-order statistics, and do not easily segment.

2.3 Statistics of textons
However, researchers soon found counterexamples to the Julesz conjecture (Caelli and Julesz 1978; Caelli et al 1978; Julesz et al 1978; Victor and Brodie 1978). For example, the Δ texture pair (Figure 1b) is relatively easy to segment, yet the two textures have the same 2nd-order statistics. A difference in 2nd-order pixel statistics appeared neither necessary nor sufficient for texture segmentation.
Based on the importance of line orientation in texture segmentation (Beck 1966; Beck 1967; Olson and Atneave 1970), two new classes of theories emerged. The first suggested that texture segmentation was mediated not by 2nd-order pixel statistics, but rather by 1st-order statistics of basic stimulus features such as orientation and size (Beck, Prazdny, and Rosenfeld, 1983). Here “1st-order” refers to histograms of, e.g., orientation, instead of pixel values.

But what of the Δ→ texture pair? By construction, it contained no difference in the 1st-order statistics of line orientation. However, notably triangles are closed shapes, whereas arrows are not. Perhaps emergent features (Pomerantz & Cragin, this volume), like closure, also matter in texture segmentation. Other iso-2nd order pairs hinted at the relevance of additional higher-level features, dubbed textons. Texton theory proposes that segmentation depends upon 1st-order statistics not only of basic features like orientation, but also of textons such as curvature, line endpoints, and junctions (Julesz 1981; Bergen and Julesz 1983).

While intuitive on the surface, this explanation was somewhat unsatisfying. Proponents were vague about the set of textons, making the theory difficult to test or falsify. In addition, it was not obvious how to extract textons, particularly for natural images (Figure 1ef). (Though see Barth et al. (1998), for both a principled definition of a class of textons, and a way to measure them in arbitrary images.) Texton theories have typically been based on verbal descriptions of image features rather than actual measurements (Bergen and Adelson 1988). These “word models” effectively operate on “things” like “closure” and “arrow junctions” which a human experimenter has labeled (Adelson 2001).

2.4 Image processing-based models
By contrast, another class of “image computable” theories emerged. These models are based on simple image processing operations (Knutsson and Granlund 1983; Caelli 1985; Turner 1986; Bergen and Adelson 1988; Sutter et al. 1989; Fogel and Sagi 1989; Bovik et al. 1990; Malik and Perona 1990; Bergen and Landy 1991; Rosenholtz 2000). According to these theories, texture segmentation arises as an outcome of mechanisms like those known to exist in early vision.

These models have similar structure: a first linear filtering stage, followed by a non-linear operator, additional filtering, and a decision stage. They have been termed filter-rectify-filter (e.g. Dakin et al. 1999), or linear-nonlinear-linear (LNL, Landy and Graham 2004) models. Chubb and Landy (1991) dubbed the basic structure the “back-pocket model”, as it was the model many researchers would “pull out of their back pocket” to explain segmentation phenomena.

The first stage typically involves multiscale filters, both oriented and unoriented. The stage-two non-linearity might be a simple squaring, rectification, or energy computation (Knutsson and Granlund 1983; Turner 1986; Sutter et al. 1989; Bergen and Adelson 1988; Fogel and Sagi 1989; Bovik et al. 1990), contrast normalization (Landy and Bergen 1991; Rosenholtz 2000), or inhibition and excitation between neighboring channels and locations (Caelli 1985; Malik and Perona 1990). The final filtering and decision stages often act as a coarse-scale edge detector. Much effort has gone into uncovering the details of the filters and nonlinearities.

As LNL models employ oriented filters, they naturally predict segmentation of textures that differ in their component orientations. But what about results thought to require more complex texton operators? Bergen and Adelson (1988) examined segmentation of an XL texture pair like that in Figure
These textures contain the same distribution of line orientations, and Bergen and Julesz (1983) had suggested that easy segmentation might be mediated by such features as terminators and X- vs. L-junctions. Bergen and Adelson (1988) demonstrated the feasibility of a simpler solution, based on low-level mechanisms. They observed that the Xs appear smaller than the Ls, even though their component lines are the same length. Beck (1967) similarly observed that Xs and Ls have a different overall distribution of brightness when viewed out of focus. Bergen and Adelson demonstrated that if one accentuates the difference in size, by increasing the length of the Ls' bars (while compensating the bar intensities so as not to make one texture brighter than the other), segmentation gets easier. Decrease the length of the Ls' bars, and segmentation becomes quite difficult. Furthermore, they showed that in the original stimulus, a simple size-tuned mechanism – center-surround filtering followed by full-wave rectification – responds more strongly to one texture than the other. Even though our visual systems can ultimately identify nameable features like terminators and junctions, those features may not underlie texture segmentation, which may involve lower-level mechanisms.

The LNL models naturally lend themselves to implementation. Nearly all the models cited here (Section 2.4) were implemented at least up to the decision stage. They operate on arbitrary images. Implementation makes these models testable and falsifiable, in stark contrast to word models operating on labeled “things” like micropatterns and their features. Furthermore, the LNL models have performed reasonably well. Malik and Perona’s (1990) model, one of the most fully specified and successful, made testable predictions of segmentation difficulty for a number of pairs of micropattern textures. They found strong agreement between their model’s predictions and behavioral results of Kröse (1986) and Gurnsey and Browse (1987). They also produced meaningful results on a complex piece of abstract art. Image computable models naturally make testable predictions about the effects of texture density (Rubenstein and Sagi 1996) alignment, and sign of contrast (Graham et al. 1992; Beck et al. 1987), for which word models inherently have trouble making predictions.

### 2.5 Bringing together statistical and image processing-based models

Is texture segmentation, then, a mere artifact of early visual processing, rather than a meaningful indicator of statistical differences between textures? The visual system should identify boundaries in an intelligent way, not leave their detection to the caprices of early vision. Making intelligent decisions in the face of uncertainty is the realm of statistics. Furthermore, statistical models seem appropriate due to the statistical nature of textures.

Statistical and image processing-based theories are not mutually exclusive. Arguably the first filtering stage in LNL models extracts basic features, and the later filtering stage computes a sort of average. Perhaps thinking in terms of intelligent decisions can clarify the role of unknown parameters in the LNL models, better specify the decision process, and lend intuitions about which textures segment.

If the mean orientations of two textures differ, should we necessarily perceive a boundary? From a decision-theory point of view this would be unwise; a small difference in mean might occur by chance. Perhaps textures segment if their 1st-order feature statistics are significantly different (Voorhees and Poggio 1988; Puzicha et al. 1997; Rosenholtz 2000). Significant difference takes into account the variability of the textures; two homogeneous textures with mean orientations differing by 30 degrees may segment, while two heterogeneous textures with the same difference in mean may not. Experimental results confirm that texture segmentation shows this dependence upon texture variability.
Observers can also segment two textures differing significantly in the variance of their orientations. However, observers are poor at segmenting two textures with the same mean and variance, when one is unimodal and the other bimodal (Rosenholtz 2000). It seems that observers do not use the full 1st-order statistics of orientation.

These results point to the following model of texture segmentation (Rosenholtz 2000). The observer collects $n$ noisy feature estimates from each side of a hypothesized edge. The number of samples is limited, as texture segmentation involves local rather than global statistics (Nothdurft 1991). If the two sets of samples differ significantly, with some confidence, $\alpha$, then the observer sees a boundary. Rosenholtz (2000) tests for a significant difference in mean orientation, mean contrast, orientation variance, and contrast variance.

The model can be implemented using biologically plausible image processing operations. Though the theoretical development came from thinking about statistical tests on discrete samples, the model extracts no “things” like line elements or texels. Rather it operates on continuous “stuff” (Adelson 2001). The model has three fairly intuitive free parameters, all of which can be determined by fitting behavioral data. Two internal noise parameters capture human contrast and orientation discriminability. The last parameter specifies the radius of the region over which measurements are pooled to compute the necessary summary statistics (mean, variance, etc.).

Human performance segmenting orientation-defined textures is well fit by the model (Rosenholtz 2000). The model also predicts the rank ordering of segmentation strength for micropattern texture pairs (TL, +T, Δ→, and L+) found by Gurnsey and Browse (1987). Furthermore, Hindi Attar et al. (2007) related the salience of a texture boundary to the rate of filling-in of the central texture in stabilized images. They found that the model predicted many of the asymmetries found in filling-in.

The visual system may do something intelligent, like a statistical test (Voorhees and Poggio 1988; Puzicha et al. 1997; Rosenholtz 2000), or Bayesian inference (Lee 1995, Feldman, this volume, Chapter 54 on Bayesian models), when detecting texture boundaries within an image. These decisions can be implemented using biologically plausible image processing operations, thus bringing together image processing-based and statistical models of texture segmentation.

3. Texture perception more broadly

Decisions based upon a few summary statistics do a surprisingly good job of predicting existing texture segmentation phenomena. Are these few statistics all that is required for texture perception more broadly? This seems unlikely. First, they perhaps do not even suffice to explain texture segmentation. Simple contrast energy has probably worked in place of more complex features only because we have tested a very limited a set of textures (Barth et al. 1998).

Second, consider Figure 1a-d. The mean and variance of contrast and orientation do little to capture the appearance of the component texels, yet we have a rich percept of their shapes and arrangement. What measurements, then, might human vision use to represent textures?
Much of the early work in texture classification and discrimination came from computer vision. It aimed at distinguishing between textured regions in satellite imagery, microscopy, and medical imagery. As with texture segmentation, early research pinpointed 2nd-order statistics, particularly the power spectrum, as a possible representation (Bajcsy 1973). Researchers also explored Markov Random Field representations more broadly. For practical applications, power spectrum and related measures worked reasonably well. (For a review, see Haralick 1979, and Wechsler 1980.)

However, the power spectrum cannot predict texture segmentation, and texture appearance likely requires more information rather than less. Furthermore, texture classification provides a weak test. Performance is highly dependent upon both the diversity of textures in the dataset and the choice of texture categories. A texture analysis/synthesis method better enables us to get a sense of the information encoded by a given representation (Tomita et al. 1982; Portilla and Simoncelli 2000). Texture analysis/synthesis techniques measure a descriptor for a texture, and then generate new samples of texture which share the same descriptor. Rather than simply synthesizing a texture with given properties, they can measure those properties from an arbitrary input texture. The “analysis” stage makes the techniques applicable to a far broader array of textures. Most of the progress in developing models of human texture representation has been made using texture analysis/synthesis strategies.

One can easily get a sense of the information encoded by the power spectrum by generating a new image with the same Fourier transform magnitude, but random phase. This representation is clearly inadequate to capture the appearance (Figure 2). The synthesized texture in Figure 2b looks like filtered noise (because it is), rather than like the peas in Figure 2a. The synthesized texture has none of the edges, contours, or other locally oriented structures of a natural image. Natural images are highly non-Gaussian (Zetzsche et al 1993). The responses of oriented bandpass filters applied to natural scenes are kurtotic (sparse) and highly dependent; these statistics cannot be captured by the power spectrum alone, and are responsible for important aspects of the appearance of natural images (Simoncelli and Olshausen 2001).

![Figure 2. Comparison of the information encoded in different texture descriptors. (a) Original peas image; (b) Texture synthesized to have the same power spectrum as (a), but random phase. This representation cannot capture the structures visible in many natural and artificial textures, though it performs adequately for some textures such as the left side of Figure 1e. (c) Marginal statistics of multiscale, oriented and non-oriented filter banks better capture the nature of edges in natural images (Heeger and Bergen](image-url)
Joint statistics work even better at capturing structure (Portilla and Simoncelli 2000).

Due to limitations of the power spectrum and related measures, researchers feared that statistical descriptors could not adequately capture the appearance of textures formed of discrete elements, or containing complex structures (Tomita et al. 1982). Some researchers abandoned purely statistical descriptors in favor of more “structural” approaches, which described texture in terms of discrete texels and their placement rule (Tomita et al. 1982; Zucker 1976; Haralick 1979). Implicitly, structural approaches assume that texture processing occurs at later stages of vision, “a cognitive rather than a perceptual approach” (Wechsler 1980). Some researchers suggested choosing between statistical and structural approaches, depending upon the kind of texture (Zucker 1976; Haralick 1979).

Structural models were less than successful, largely due to difficulty extracting texels. This worked better when texels were allowed to consist of arbitrary image regions, rather than correspond to recognizable “things” (e.g. Leung and Malik 1996).

The parallels to texture segmentation should be obvious: researchers rightly skeptical about the power of simple statistical models abandoned them in favor of models operating on discrete “things”. As with texture segmentation, the lack of faith in statistical models proved unfounded. Sufficiently rich statistical models can capture a lot of structure. Demonstrating this requires more complex texture synthesis methodologies to find samples of texture with the same statistics. A number of texture synthesis techniques have been developed, with a range of proposed descriptors.

Heeger and Bergen’s (1995) descriptor, motivated by the success of the LNL segmentation models, consists of marginal (i.e. 1st-order) statistics of the outputs of multiscale filters, both oriented and unoriented. Their algorithm synthesizes new samples of texture by beginning with an arbitrary image “seed” – often a sample of random noise, though this is not required – and iteratively applying constraints derived from the measured statistics. After a number of iterations, the result is a new image with (approximately) the same 1st-order statistics as the original. Figure 2c shows an example. Their descriptor captures significantly more structure than the power spectrum; enough to reproduce the general size of the peas and their dimples. It still does not quite get the edges right, and misrepresents larger-scale structures.

Portilla and Simoncelli (2000) extended the Heeger/Bergen methodology, and included in their texture descriptor the joint (2nd-order) statistics of responses of multiscale V1-like simple and complex “cells”. Figure 2d shows an example synthesis. This representation captures much of the perceived structure, even in micropattern textures (Portilla and Simoncelli 2000; Balas 2006), though it is not perfect. Some non-parametric synthesis techniques have performed better at producing new textures which look like the original (e.g. Efros and Leung 1999). However, these techniques use a texture descriptor which is essentially the entire original image. It is unclear how biologically plausible such a representation might be, or what the success of such techniques teach us about human texture perception.

Portilla and Simoncelli (2000), then, remains a state-of-the-art parametric texture model. This does not imply that its measurements are literally those made by the visual system, though they are certainly biologically plausible. A “rotation” of the texture space would maintain the same information while
changing the representation dramatically. Furthermore, a sufficiently rich set of 1st-order statistics can encode the same information as higher-order statistics (Zhu et al 1996). However, the success of Portilla and Simoncelli’s model demonstrates that a rich and high-dimensional set of image statistics comes close to capturing the information preserved and lost in visual representation of a texture.

4. Texture perception is not just for textures

Researchers have long studied texture perception in the hope that it would lend insight into vision more generally. Texture segmentation, rather than merely informing us about perceptual organization, might uncover the basic features available preattentively (Treisman 1985), or the nature of early nonlinearities in visual processing (Malik and Perona 1990; Graham et al. 1992; Landy and Graham 2004). However, common wisdom assumed that after the measurement of basic features, texture and object perception mechanisms diverged (Cant and Goodale 2007). Similarly, work in computer vision assumed separate processing for texture vs. objects.

More recent work blurs the distinction between texture and object processing. Modern computer vision treats them much more similarly. Recent human vision research demonstrates that “texture processing” operations underlie vision more generally. The field’s previous successes in understanding texture perception may elucidate visual processing for a broad array of tasks.

4.1 Peripheral crowding

Texture processing mechanisms have been associated with visual search (Treisman 1985) and set perception (Chong and Treisman 2003). One can argue that texture statistics naturally inform these tasks. Evidence of more general texture processing in vision has come from the study of peripheral vision, in particular visual crowding.

Peripheral vision is substantially worse than foveal vision. For instance, the eye trades off sparse sampling over a wide area in the periphery for sharp, high resolution vision over a narrow fovea. If we need finer detail, we move our eyes to bring the fovea to the desired location.

The phenomenon of visual crowding2 illustrates that loss of information in the periphery is not merely due to reduced acuity. A target such as the letter ‘A’ is easily identified when presented in the periphery on its own, but becomes difficult to recognize when flanked too closely by other stimuli, as in the string of letters, ‘BOARD’. An observer might see these crowded letters in the wrong order, perhaps confusing the word with ‘BORAD’. They might not see an ‘A’ at all, or might see strange letter-like shapes made up of a mixture of parts from several letters (Lettvin 1976).

2 “Crowding” is used inconsistently and confusingly in the field, sometimes as a transitive verb (“the flankers crowd the target”), sometimes as a mechanism, and sometimes as the experimental outcome in which recognizing a target is impaired in the presence of nearby flankers. This chapter predominantly follows the last definition, though in describing stimuli sometimes refers to the lay “at lot of stuff in a small space.”
Crowding occurs with a broad range of stimuli (see Pelli and Tillman 2008 for a review). However, not all flankers are equal. When the target and flankers are dissimilar or less grouped together, target recognition is easier (Andriessen and Bouma 1976; Kooi et al 1994; Saarela et al. 2009). Strong grouping among the flankers can also make recognition easier (Livne and Sagi 2007; Sayim et al 2010; Manassi et al. 2012). Furthermore, crowding need not involve discrete “target” and “flankers”; Martelli et al. (2005) argue that “self-crowding” occurs in peripheral perception of complex objects and scenes.

4.2 Texture processing in peripheral vision?
The percept of a crowded letter array contains sharp, letter-like forms, yet they seem lost in a jumble, as if each letter’s features (e.g., vertical bars and rounded curves) have come untethered and been incorrectly bound to the features of neighboring letters (Pelli et al. 2004). Researchers have associated the phenomena of crowding with the “distorted vision” of strabismic amblyopia (Hess 1982). Lettvin (1976) observed that an isolated letter in the periphery seems to have characteristics which the same letter, flanked, does not. The crowded letter “only seems to have a ‘statistical’ existence.” In line with these subjective impressions, researchers have proposed that crowding phenomena result from “forced texture processing,” involving excessive feature integration (Pelli et al. 2004), or compulsory averaging (Parkes et al. 2001) over each local pooling region. Pooling region size grows linearly with eccentricity, i.e. with distance to the point of fixation (Bouma 1970).

Assume for the sake of argument – following Occam’s razor – that the peripheral mechanisms underlying crowding operate all the time, by default; no mechanism perversely “switches on” to thwart our recognition of flanked objects. This Default Processing assumption has profound implications for vision. Peripheral vision is hugely important; very little processing truly occurs in the fovea. One can easily recognize the cat in Figure 3, when fixating on the “+”. Yet the cat may extend a number of degrees beyond the fovea. Could object recognition, perceptual organization, scene recognition, face recognition, navigation, and guidance of eye movements all share an early, local texture processing mechanism? Is it that “texture is primitive and textures combine to produce forms” (Lettvin 1976)? This seems antithetical to ideas of different processing for textures and objects. Prior to 2000, it would have seemed surprising to use a texture-like representation for more general visual tasks.
Figure 3. Original images (a,c) and images synthesized to have approximately the same local summary statistics (b,d). Intended (and model) fixation on the “+”. The cat can clearly be recognized while fixating, even though much of the object falls outside the fovea. The summary statistics contain sufficient information to capture much of its appearance (b). Similarly, the summary statistics contain sufficient information to
recognize the gist of the scene (d), though perhaps not to correctly assess its details. (e) A patch of search display, containing a tilted target and vertical distractors. (f) The summary statistics (here, in a single pooling region) are sufficient to decipher the approximate number of items, much about their appearance, and the presence of the target. (g) A target-absent patch from search for a white vertical among black vertical and white horizontal. (h) The summary statistics are ambiguous about the presence of a white vertical, perhaps leading to perception of illusory conjunctions. (c-h) originally appeared in (Rosenholtz, Huang, et al. 2012).

However, several state-of-the-art computer vision techniques operate upon local texture-like image descriptors, even when performing object and scene recognition. The image descriptors include local histograms of gradient directions, and local mean response to oriented multi-scale filters, among others (Bosch et al 2006, 2007; Dalal and Triggs, 2005; Oliva and Torralba 2006; Tola et al. 2010; Fei-Fei and Perona 2005). Such texture descriptors have proven effective for detection of humans in natural environments (Dalal and Triggs, 2005), object recognition in natural scenes (Bosch et al, 2007; Mutch and Lowe, 2008; Zhu, Bichot, and Chen, 2011), scene classification (Oliva and Torralba 2001; Renninger and Malik 2004; Fei-Fei and Perona 2005), wide-baseline stereo (Tola et al. 2010), gender discrimination (Wang, Yau, and Sung, 2010), and face recognition (Velardo and Dugelay, 2010). These results represent only a handful of hundreds of recent computer vision papers utilizing similar methods.

Suppose we take literally the idea that peripheral vision involves early local texture processing. The key questions are whether on the one hand, humans make the sorts of errors one would expect, and on the other hand whether texture processing preserves enough information to explain the successes of vision, such as object and scene recognition.

A local texture representation predicts vision would be locally ambiguous in terms of the phase and location of features, as texture statistics contains such ambiguities. Do we see evidence in vision? In fact, we do. Observers have difficulty distinguishing 180 degree phase differences in compound sine wave gratings in the periphery (Bennett and Banks 1991; Rentschler and Treutwein 1985) and show marked position uncertainty in a bisection task (Levi and Klein 1986). Furthermore, such ambiguities appear to exist during object and scene processing, though we rarely have the opportunity to be aware of them. Peripheral vision tolerates considerable image variation without giving us much sense that something is wrong (Freeman and Simoncelli 2011; Koenderink et al. 2012). Koenderink et al. (2012) apply a spatial warping to an ordinary image. It is surprisingly difficult to tell that anything is wrong, unless one fixates near the image. (See http://i-perception.perceptionweb.com/fulltext/i03/i0490sas.)

To go beyond qualitative evidence, we need a concrete proposal for what “texture processing” means. This chapter has reviewed much of the relevant work. Texture appearance models aim to understand texture processing in general, whereas segmentation models attempt only to predict grouping. Our current best guess as to a model of texture appearance is that of Portilla and Simoncelli (2000). Perhaps the visual system computes something like 2\textsuperscript{nd}-order statistics of the responses of V1-like cells, over each local pooling region. We call this the Texture Tiling Model. This proposal (Balas et al. 2009; Freeman and Simoncelli 2011) is not so different from standard object recognition models, in which later stages compute more complex features by measuring co-occurrences of features from the previous
layer (Fukushima 1980; Riesenhuber and Poggio 1999). Second-order correlations are essentially co-occurrences pooled over a substantially larger area.


Visual search employs wide field-of-view, crowded displays. Is the difference between easy and difficult search due to local texture processing? We can utilize texture synthesis techniques to visualize the local information available (Figure 3). When target and distractor bars differ significantly in orientation, the statistics are sufficient to identify a crowded peripheral target. The model predicts easy “popout” search. The model also predicts the phenomenon of illusory conjunctions, and other classic search results (Rosenholtz, Huang, et al. 2012; Rosenholtz, Raj, et al. 2012). Characterizing visual search as limited by peripheral processing represents a significant departure from earlier interpretations which attributed performance to the limits of processing in the absence of covert attention (Treisman 1985).

Under the Default Processing assumption, we must also ask whether texture processing might underlie normal object and scene recognition. We synthesized an image to have the same local summary statistics as the original (Rosenholtz 2011; Rosenholtz, Huang, et al. 2012; see also Freeman and Simoncelli 2011). A fixed object (Figure 3b) is clearly recognizable; it is quite well encoded by this representation. Glancing at a scene (Figure 3d), much information is available to deduce the gist and guide eye movements; however, precise details are lost, perhaps leading to change blindness (Oliva and Torralba 2006; Freeman and Simoncelli 2011; Rosenholtz, Huang, et al. 2012).

These results and demos indicate the power of the Texture Tiling Model. It is image computable, and can make testable predictions for arbitrary stimuli. It predicts on the one hand difficulties of vision, such as crowded object recognition and hard visual search, while plausibly supporting normal object and scene recognition.

### 4.3 Parallels between alternative models of crowding and less successful texture models

It is instructive to consider alternative models of crowding, and their parallels to previous work on texture perception. A number of crowding experiments have been designed to test an overly simple texture processing model. In this “simple pooling” or “faulty-integration” model, each pooling region yields the mean of some (often unspecified) feature. To a first approximation, this model predicts worse performance the more one fills up the pooling region with irrelevant flankers, as doing so reduces the informativeness of the mean. This impoverished model cannot explain improved performance with larger flankers (Levi and Carney 2009; Manassi et al. 2012), or when flankers group with one another (Saarela et al. 2009; Manassi et al. 2012).

Partially in response to failures of the simple pooling model, researchers have suggested that some grouping might occur prior to the mechanisms underlying crowding (Saarela et al. 2009). More generally, the field tends to describe crowding mechanisms as operating on “things”: Levi and Carney (2009) suggested that a key determinant of whether crowding occurs is the distance between target and flanker centroids. Averaging might operate on discrete features of objects within the pooling region (Parkes et al. 2001; Greenwood et al. 2009; Pöder and Wagemans 2007; Greenwood et al. 2012), and/or
localization of those discrete features might be poor (Strasburger 2005; van den Berg et al. 2012). Some crowding effects seem to depend upon target/flanker identities rather than their features (Louie et al. 2007; Dakin et al. 2010), suggesting that they may be due to later, object-level mechanisms. Though as Dakin et al. (2010) demonstrate, these apparently “object-centered” effects can be explained by lower-level mechanisms.

This sketch of alternative models should sound familiar. That crowding mechanisms might act after early operations have split the input into local groups or objects should have obvious parallels to theories of texture perception. Once again, a too-simple “stuff” model has been rejected in favor of models which operate on “things”. These models, typically word models, do not easily make testable predictions for novel stimuli.

4.4 The power of pooling in high dimensions
A “simple pooling model” bears little resemblance to successful texture descriptors. Texture perception requires a high dimensional representation. The Portilla and Simoncelli (2000) texture model computes 700-1000 image statistics per texture (depending upon choice of parameters). (The Texture Tiling Model computes this many statistics per local pooling region.) The “forced texture perception” presumed to underlie crowding must also be high dimensional – after all, it must at the very least support perception of actual textures.

Unfortunately it is difficult in general to get intuitions about behavior of high-dimensional models. Low-dimensional models do not simply scale up to higher dimensions. A single mean feature value captures little information about a stimulus. Additional statistics provide an increasingly good representation of the original patch. Stuff-models, if sufficiently rich, can in fact capture a great deal of information about the visual input.

How well a stimulus can be encoded depends upon its complexity relative to the representation. Flanker grouping can theoretically simplify the stimulus, leading to better representation and perhaps better performance. In some cases the information preserved is insufficient to perform a given task, and in common parlance the stimulus is “crowded.” In other cases, the information is sufficient for the task, predicting the “relief from crowding” accompanying, for example, a dissimilar target and flankers (e.g. Rosenholtz, Raj, et al. 2012 and Figure 3e-h).

A high-dimensional representation can also preserve the information necessary to individuate “things”. For instance, it can capture the approximate number of discrete objects in Figure 3eg. In fact, one can represent an arbitrary amount of structure in the input by varying the size of the regions over which statistics are computed (Koenderink and van Doorn 2000), and the set of statistics. The structural/statistical distinction is not a dichotomy, but rather a continuum.

The mechanisms underlying crowding may be “later” than texture perception mechanisms, and operate on precomputed groups or “things”. However, just because we often recognize “things” in our stimuli, as a result of the full visual-cognitive machinery, does not mean that our visual systems operate upon those things to perform a given task. One should not underestimate the power of high-dimensional models which operate on continuous “stuff”. In texture perception, such models have explained results for a wider variety of stimuli, and with arguably simpler mechanisms.
5. Conclusions

In the last several decades, much progress has been made toward better understanding the mechanisms underlying texture segmentation, classification, and appearance. There exists a rich body of work on texture segmentation, both behavioral experiments and modeling. Many results can be explained by intelligent decisions based on some fairly simple image statistics. Researchers have also developed powerful models of texture appearance. More recent work demonstrates that similar texture-processing mechanisms may account for the phenomena of visual crowding. The details remain to be worked out, but if true, the visual system may employ local texture processing throughout the visual field. This predicts that, rather than being relegated to a narrow set of tasks and stimuli, texture processing underlies visual processing in general, supporting such diverse tasks as visual search, object and scene recognition.
6. References


