

Region grouping in natural foliage scenes: Image statistics and human performance

Almon D. Ing

Department of Psychology and Center for Perceptual Systems,
University of Texas at Austin, Austin, USA



J. Anthony Wilson

Department of Psychology and Center for Perceptual Systems,
University of Texas at Austin, Austin, USA



Wilson S. Geisler

Department of Psychology and Center for Perceptual Systems,
University of Texas at Austin, Austin, USA



This study investigated the mechanisms of grouping and segregation in natural scenes of close-up foliage, an important class of scenes for human and non-human primates. Close-up foliage images were collected with a digital camera calibrated to match the responses of human *L*, *M*, and *S* cones at each pixel. The images were used to construct a database of hand-segmented leaves and branches that correctly localizes the image region subtended by each object. We considered a task where a visual system is presented with two image patches and is asked to assign a category label (either *same* or *different*) depending on whether the patches appear to lie on the *same* surface or *different* surfaces. We estimated several approximately ideal classifiers for the task, each of which used a unique set of image properties. Of the image properties considered, we found that ideal classifiers rely primarily on the difference in average intensity and color between patches, and secondarily on the differences in the contrasts between patches. In psychophysical experiments, human performance mirrored the trends predicted by the ideal classifiers. In an initial phase without corrective feedback, human accuracy was slightly below ideal. After practice with feedback, human accuracy was approximately ideal.

Keywords: structure of natural images, color vision, spatial vision, perceptual organization

Citation: Ing, A. D., Wilson, J. A., & Geisler, W. S. (2010). Region grouping in natural foliage scenes: Image statistics and human performance. *Journal of Vision*, 10(4):10, 1–19, <http://journalofvision.org/10/4/10/>, doi:10.1167/10.4.10.

Introduction

Humans are endowed with a remarkable ability to correctly interpret the two-dimensional images formed on their retinas. They can reliably assign retinal locations to distinct physical surfaces, identify whether image contours are surface boundaries, lighting boundaries, or internal surface markings, identify the material composition of surfaces, determine the shapes and distances of surfaces, and group surfaces into distinct objects. These abilities reflect the visual system's implicit understanding of the relationships between the physical properties of the natural environment and the properties of retinal images. Given this realization and the recent increases in available computer power, there has been a growing effort to measure and understand natural scene statistics (for reviews, see Geisler, 2008; Reinagel, 2001; Simoncelli, 2003; Simoncelli & Olshausen, 2001).

The goal in many natural visual tasks is to make accurate inferences about properties of the environment ω (distal stimuli) from properties of the retinal image s (proximal stimuli), where ω and s represent properties in the two domains. The key statistic needed for making such inferences is the joint probability distribution of the

environment and image properties, $p(\omega, s)$, which gives directly the posterior probability of environment properties given observed image properties, $p(\omega|s)$. If these posterior probabilities and the costs and benefits of the different possible environment-behavior outcomes are known, then it is possible in principle to derive the Bayesian ideal observer for the task and to determine how well that ideal observer performs (e.g., see Geisler, 2008; Geisler & Diehl, 2003; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996). The computations carried out by the ideal observer provide insight into what *should* be computed by the visual system when performing the task, and hence can suggest principled hypotheses for perceptual mechanisms. The performance of the ideal observer quantifies the potential usefulness of the particular image properties under consideration and provides an appropriate benchmark against which to compare human performance.

In the present study, we measured natural scene statistics, derived approximate ideal observers, and measured human performance in a simple perceptual grouping task that is relevant to the more general task of segmentation and grouping in natural foliage environments. Specifically we consider a simple *patch grouping task* where the observer is given two equal size image

patches sampled from a natural foliage image at some spatial separation and must decide whether they belong to the *same* or *different* physical surfaces.

To measure the relevant natural scene statistics we collected a diverse set of calibrated images of close-up foliage which were then hand segmented by human observers. We chose to analyze close-up foliage because (i) foliage images are a major component of the environment and the dominant component in the natural environment of the macaque monkey (the primary animal model for human vision), (ii) many perceptual-motor tasks performed by the macaque involve interacting with close-up foliage (e.g., picking or moving leaves, grabbing branches or fruit), (iii) close-up foliage images are relatively easy to hand segment, (iv) the statistical properties of distance foliage should be derivable in part from those of close-up foliage, and (v) solutions to the complex problem of foliage segmentation may generalize robustly to other image segmentation problems.

A number of previous studies have used hand-segmented images for measuring natural scene statistics (Balboa & Grzywacz, 2000; Brunswik & Kamiya, 1953; Elder & Goldberg, 2002; Fowlkes, Martin, & Malik, 2007; Geisler, Perry, Super, & Gallogly, 2001; Konishi, Yuille, Coughlan, & Zhu, 2003; Martin, Fowlkes, & Malik, 2004; Torralba, 2009). In the present study, a central assumption is that the hand segmentation provides the approximate “ground-truth” linkage between the image and the physical environment. In other words, we assumed that when a human observer segments a leaf or branch from the background, the segmented image boundary corresponds to the true surface boundary in the environment. This is a critical assumption because knowing the approximate ground truth is necessary for estimating the joint-probability distribution $p(\omega, s)$. The obvious pitfall of the hand-segmentation approach is that there may be conditions where human segmentations are not reliable enough to approximate ground truth, although this is unlikely to be a problem in our case (see later).

The task and image statistics considered here were motivated in part by Fine, MacLeod, and Boynton (2003), who measured the statistics of color differences between pixels at different separations within images and in different images, and then compared a model of image segmentation based on these statistics with human segmentation judgments in natural images. The specific tasks and image statistics they considered are quite different from the current study; we describe the differences between the two studies in the discussion section.

The patch grouping task is essentially a binary classification task in which the environmental state ω can assume one of two nominal values, which we represent with the nominal variable $\omega \in \{\text{same}, \text{different}\}$. There are many possible image properties (s) that could be considered, but here we focus on the mean luminance and color of each patch as well as the luminance and color contrast of each patch. The specific definitions of the

image properties are given in the methods. These local image properties were chosen because they are simple and because they allowed us to determine approximate ideal classifiers.

To examine the relationship between human performance and the measured image statistics, human performance was measured under three conditions as a function of the distance between patches. In the *texture-removed* condition, the subjects were presented with natural image patches that were modified to preserve only the difference in mean luminance and color. Comparison of human and ideal performance in this case allowed us to determine how efficiently humans use this information. In the *full* condition, the subjects were presented the actual natural image patches. In the *texture-only* condition, the subjects were presented natural image patches, where mean luminance and color were equated but where contrast and spatial texture information were unchanged. The *full* and *texture-only* conditions are useful for determining whether humans use (in this task) image properties in addition to those we measured.

It is important to keep in mind that our aim is to measure scene statistics and performance for selected local image properties. If humans are given the entire image they make essentially no mistakes and indeed we are using this fact to obtain the approximate ground truth for our tasks. However, when humans are given just a small image patch they make mistakes (Geisler & Perry, 2009; McDermott, 2004), and they make even more mistakes if they are given only the image properties we measured. The central questions we are asking are these: How much and what information is available in the local image patches for performing the patch grouping task? And, how efficient are humans at using that information?

Methods

Natural scene statistics methods

Camera calibration

A 36-bit-per-pixel Kodak DCS720x digital CCD camera (12 bits per color sensor) was used to capture the close-up images of foliage. The camera was calibrated to give the approximate responses in the human long (L), middle (M) and short (S) wavelength cones at each pixel location. In the first step of the calibration, we verified that the camera responded linearly over its dynamic range and that the f-stop and shutter speed controls functioned accurately.

In the second step, we measured the sensitivity of the camera sensors at each wavelength. A monochromator was used to cast narrow-band light onto a white reflectance standard where test wavelengths ranged from 394 nm to 720 nm. For each test wavelength, we measured both the camera’s response to the standard and

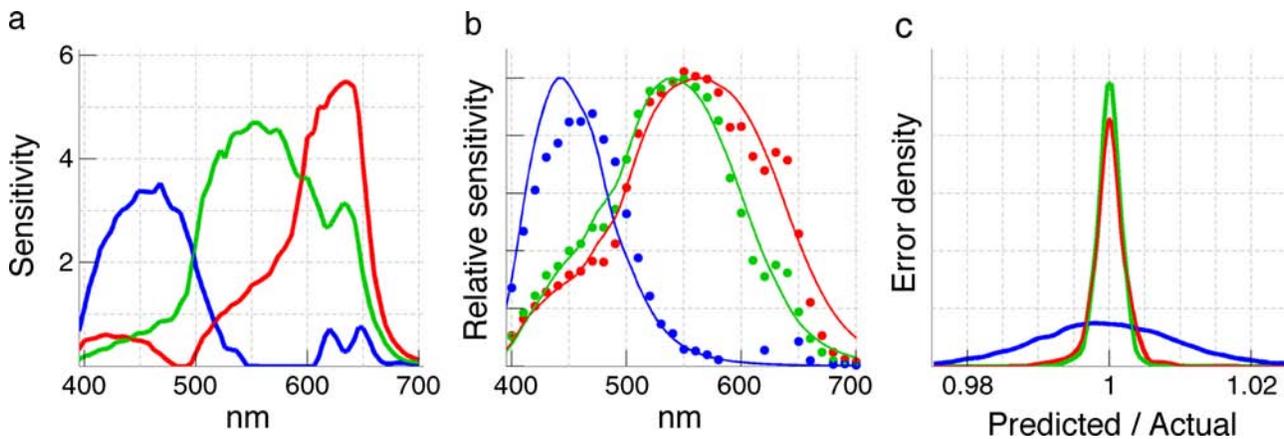


Figure 1. **a.** The sensitivity of the camera's red, green, and blue sensors is shown in units of response per 10^{15} q/s/ster/m²/nm. For example, the red sensor gives a value of about 5 in response to a 650 nm light with a radiance of 10^{15} q/s/ster/m². **b.** The sensitivity of *L* (red), *M* (green), and *S* (blue) cones (in quantal units) based on the CIE 10-deg color matching functions adjusted to 2-deg according to Stockman et al. (1993) are plotted as thin curves. The camera estimates are plotted as dots. **c.** Three histograms showing the accuracy of the camera's estimation of the *L* (red), *M* (green), and *S* (blue) cones for simulated natural radiance spectra.

the radiance spectrum of the standard (with a Photo-Research PR704 spectrophotometer). The sequence of measurements was performed twice to reduce errors. From these measurements, we estimated the spectral sensitivities of the camera's red (*R*), green (*G*), and blue (*B*) sensors at every wavelength (see Figure 1a).

In the third step, the camera's spectral sensitivity functions were linearly transformed to mimic the spectral sensitivities of *L*, *M*, and *S* cones defined as the Stockman, MacLeod, and Johnson (1993) 2-deg cone fundamentals based on CIE 10-deg color matching functions adjusted to 2-deg (solid curves in Figure 1b). The optimal transformation matrix was estimated by minimizing the squared error between predicted and actual $\log L$, $\log M$, and $\log S$ responses to natural radiance spectra. Each natural radiance spectrum was simulated by multiplying a randomly selected natural reflectance spectrum (Krinov, 1947) with a randomly selected natural irradiance spectrum (Dicarlo & Wandell, 2000). The histograms of prediction errors is shown in Figure 1c. As can be seen the errors are generally less than a half percent for the *L* and *M* responses and less than 1% for the *S* responses. The camera's effective *L*, *M*, and *S* sensitivity functions are given by the dots in Figure 1b. We also confirmed the accuracy of the camera's calibration by comparing camera and spectrophotometer measurements of the *L*, *M*, and *S* responses for the test patches of a MacBeth color checker illuminated by a tungsten source.

Image collection

Most of the images were captured at Zilker Botanical Gardens in Austin, Texas because of the abundant variety of plant species. The remaining images were captured in rural regions of Colorado Rocky Mountains. About one half of the images were captured on cloudy days, the

remainder on sunny days. Most images were captured from the standing or crouching position with the goal of capturing close-up foliage. Most images did not include patches of sky. On sunny days, the goal was to capture close-up foliage with strong shadow boundaries. As a result, the database contains a varied sample of foliage captured under a variety of natural illumination conditions.

Segmentation

Hand-segmentation was performed by paid undergraduate students at the University of Texas at Austin, using custom software. To obtain representative statistics it was necessary to segment a large number of images. However, careful segmentation is time consuming and thus it was not practical to perform a full segmentation of each image. Thus, prior to performing the segmentation, one of the authors (ADI) defined a circular region of interest within each image, and the undergraduates were instructed to segment all objects inside or touching each region. The regions were selected to contain a large number of occlusions or shadow boundaries. By selecting a region from each image, we were able to obtain a dense segmentation for a relatively large sample of foliage images.

Figure 2 shows an example segmented image. The segmentation involved creating polygons that defined the boundary of each object's visible surface regions. Because of occlusions, multiple polygons were sometimes required to segment an object. Because an occluding object shares a common edge with the occluded object, vertices could be shared by immediately adjacent polygons.

Next the leaves were categorized for quality. If anything about a segmentation boundary was ambiguous or uncertain, the undergraduate was instructed to label it as *low quality*. The remaining boundaries were labeled



Figure 2. This example of a hand-segmented image demonstrates how polygons (colored blue with white borders) were used to segment leaves from images. All leaves were segmented that intersected or fell within a circular regions of interest (shown in orange).

high quality. As a final quality control step, one of the authors (ADI) meticulously fixed all obvious errors in the quality labeling (this was a small percentage). All low quality leaves were then discarded, leaving 1,645 high quality segmented leaf boundaries. Although this restriction may limit the generality of the results to some extent, it guarantees that the segmentations closely approximate ground truth. Furthermore, even with this restriction, the regions of interest were densely segmented. Thus, we believe the statistics reported here are highly relevant to natural tasks in the world of close-up foliage (and perhaps in other environments as well). Figure 3 shows the distribution of the number of polygons used to segment a leaf and distribution of leaf diameters (square root of the total number of pixels in the leaf) in the entire database. The quality of the segmentations is best appreciated by visiting the website <http://www.cps.utexas.edu/FoliageDB>, which contains images that illustrate the segmentations as well as a database containing the images and segmented objects.

Color space for image properties

Ruderman, Cronin, and Chiao (1998) analyzed hyperspectral camera images of foliage rich scenes and found that the distribution of the logarithms of the L , M , and S cone responses is approximately Gaussian. They determined the three principal axes of the three-dimensional Gaussian distribution using Principal Component Analysis (PCA). These axes are convenient because the marginal

distributions along each axis amount to a complete description of the full three-dimensional probability distribution of cone responses. We applied PCA to the distribution of $\log L$, $\log M$ and $\log S$ responses measured for our image set and we obtained the same eigenvectors reported by Ruderman et al. (see Table 1). Therefore we chose to define our intensity and contrast image properties in an $l\alpha\beta$ space like Ruderman et al. given by the following transformations:

$$\begin{aligned} l &= \left(\frac{\log L - \langle \log L \rangle}{\sqrt{3}} + \frac{\log M - \langle \log M \rangle}{\sqrt{3}} + \frac{\log S - \langle \log S \rangle}{\sqrt{3}} \right) \frac{1}{\sqrt{1.15}}, \\ \alpha &= \left(\frac{\log L - \langle \log L \rangle}{\sqrt{6}} + \frac{\log M - \langle \log M \rangle}{\sqrt{6}} - \frac{\log S - \langle \log S \rangle}{\sqrt{6}/2} \right) \frac{1}{\sqrt{0.0138}}, \\ \beta &= \left(\frac{\log L - \langle \log L \rangle}{\sqrt{2}} - \frac{\log M - \langle \log M \rangle}{\sqrt{2}} \right) \frac{1}{\sqrt{0.000075}}. \end{aligned} \quad (1)$$

where the brackets $\langle \rangle$ represent the operation of taking the mean. Inspection of these transformations reveals that l is a luminance-like value (which we will call “intensity”), α a blue-yellow color opponent value, and β a red-green color opponent value. The scalars on the right side of each formula convert the values to standard-deviation units (z-scores).

Patch pair properties analyzed

The properties we measured and analyzed (as a function of the distance between patches) are defined in terms of the $l\alpha\beta$ space described above. Each pair of image patches can be described in terms of twelve image properties, described below in Equations 2, 3, 4, and 5. To define these properties, let the means of patches “1” and “2” be labeled as \bar{l}_1 , \bar{l}_2 , $\bar{\alpha}_1$, $\bar{\alpha}_2$, $\bar{\beta}_1$, and $\bar{\beta}_2$; and, the standard

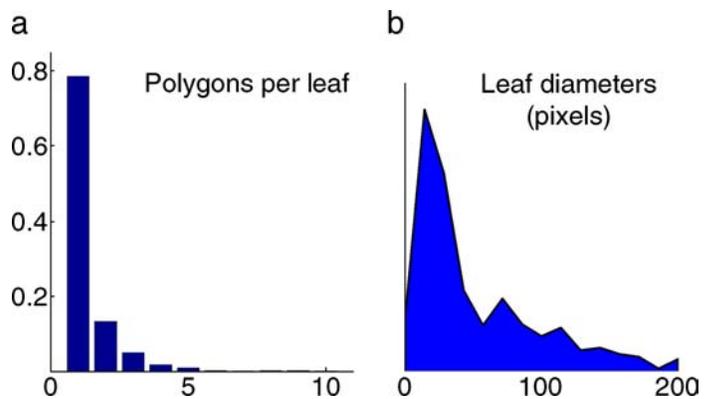


Figure 3. Summary of database of segmented leaves. **a.** Most leaves in the database (78%) were segmented using only a single polygon. **b.** A leaf’s *diameter* is defined as the square-root of the number of pixels it subtended. The leaf diameters varied widely, but most were in the range of 20–80 pixels.

Mean		Covariance Matrix			Eigenvalues			
log <i>L</i>	18.1		log <i>L</i>	log <i>M</i>	log <i>S</i>	<i>l</i>	<i>α</i>	<i>β</i>
log <i>M</i>	18.0	log <i>L</i>	0.398	0.395	0.375	1.15E+00	1.38E−02	7.52E−05
log <i>S</i>	17.3	log <i>M</i>		0.392	0.372	98.81%	1.19%	0.01%
		log <i>S</i>			0.373			
		Correlation Matrix			Eigenvectors			
		log <i>L</i>	log <i>M</i>	log <i>S</i>	<i>l</i>	<i>α</i>	<i>β</i>	
	log <i>L</i>	1.000	1.000	0.973	$1/\sqrt{3}$	$1/\sqrt{6}$	$1/\sqrt{2}$	
	log <i>M</i>		1.000	0.973	$1/\sqrt{3}$	$1/\sqrt{6}$	$-1/\sqrt{2}$	
	log <i>S</i>			1.000	$1/\sqrt{3}$	$-2/\sqrt{6}$	0	

Table 1. Shows properties of the distribution of log cone responses for a pixel randomly sampled from the entire image set. The means are given in units of $\log_{10}(q/s/sr/m^2)$.

deviations be labeled as $\sigma_{l,1}$, $\sigma_{l,2}$, $\sigma_{\alpha,1}$, $\sigma_{\alpha,2}$, $\sigma_{\beta,1}$, and $\sigma_{\beta,2}$. The labels “1” and “2” were assigned so that $\bar{l}_1 \leq \bar{l}_2$ (this assignment rule is arbitrary and has no effect on analysis results).

The first set of properties (named “ μ_3 ”) reflects the mean intensity and color of the patches:

$$\begin{aligned}\bar{l} &= \frac{1}{2}(\bar{l}_1 + \bar{l}_2), \\ \bar{\alpha} &= \frac{1}{2}(\bar{\alpha}_1 + \bar{\alpha}_2), \\ \bar{\beta} &= \frac{1}{2}(\bar{\beta}_1 + \bar{\beta}_2).\end{aligned}\quad (2)$$

(In the psychophysics experiments, it was not possible to replicate the full range of brightness observed in nature, so we also considered the set “ μ_2 ” consisting of only $\bar{\alpha}$ and $\bar{\beta}$; see later.)

Following Fine et al. (2003), the second set of properties (named “ Δ_3 ”) reflects the mean intensity and color differences:

$$\begin{aligned}\Delta l &= \bar{l}_2 - \bar{l}_1, \\ \Delta \alpha &= \bar{\alpha}_2 - \alpha_1, \\ \Delta \beta &= \bar{\beta}_2 - \bar{\beta}_1.\end{aligned}\quad (3)$$

The third set of properties (named “ δ_3 ”) reflects the contrast difference between patches:

$$\begin{aligned}\Delta \sigma_l &= \sigma_{l,2} - \sigma_{l,1}, \\ \Delta \sigma_\alpha &= \sigma_{\alpha,2} - \sigma_{\alpha,1}, \\ \Delta \sigma_\beta &= \sigma_{\beta,2} - \sigma_{\beta,1}.\end{aligned}\quad (4)$$

(Note that the standard deviation is an appropriate measure of contrast because the pixel values are in log units.)

The fourth set of image properties (named “ F_3 ”) reflects the ratio of contrasts between patches:

$$\begin{aligned}\ln F_l &= 2 \ln(\sigma_{l,2}/\sigma_{l,1}), \\ \ln F_\alpha &= 2 \ln(\sigma_{\alpha,2}/\sigma_{\alpha,1}), \\ \ln F_\beta &= 2 \ln(\sigma_{\beta,2}/\sigma_{\beta,1}).\end{aligned}\quad (5)$$

These twelve image properties are the only ones that we considered in defining the ideal classifiers.

The central properties of interest here are the mean color differences (Δ_3) and the contrast differences (δ_3) because they are intuitive and prominent in the literature. The other properties were included for completeness, because they were available to subjects in the *texture-removed* and *texture-only* stimulus conditions (and of course in the *full* condition).

Psychophysical methods

Stimuli

In the psychophysical experiments, image patches were randomly selected from the database of segmented leaves and displayed on a black background. Stimuli were displayed on a color CRT monitor (Sony GDM-FW 900) which had been calibrated using a PhotoResearch PR704 spectrophotometer and a United Detector Technologies PIN 10 photodiode. Each image pixel subtended 2.3 minutes of visual angle.

To generate a stimulus, one reference leaf was randomly sampled from the database at a time. Its *diameter* was defined as the square-root of the total number of pixels contained in the leaf. (This unit makes the results relatively invariant with viewing distance.) One circular

image patch inside the leaf was randomly selected. A second patch was selected so that its center was a distance of 1/4, 1/2, or 1 diameter from the first patch’s center. If the second patch was inside the leaf, the pair was *same*; if the second patch was outside the leaf, the pair was *different*. The diameter of the image patches was 1/5 of the reference leaf diameter. Neither patch was allowed to intersect a polygon boundary of the reference leaf.

As Table 2 summarizes, the patches were displayed in three different ways depending on the experimental condition. In the *full* condition, both patches contained all of their original properties and were essentially circular cut-outs from the image. In the *texture removed* condition, the patches were uniform, only the average intensity and color differences (ΔI , $\Delta\alpha$, and $\Delta\beta$) were preserved. In the *texture only* condition, natural contrast and spatial structure (texture) was preserved, but the difference in average intensity and color were removed by applying uniform addition, in $l\alpha\beta$ space, so that $\Delta I = 0$, $\Delta\alpha = 0$, and $\Delta\beta = 0$. In all conditions, the value \bar{I} (the average intensity of both patches) was always set to -1.35 (corresponding to approximately 4 cd/m^2) and the means of $\bar{\alpha}$ and $\bar{\beta}$ (the average color of both patches) were always preserved.

Procedure

On the first day (before the main experiment), subjects browsed through the 96 images that made up the database. They looked at each image for approximately 5 seconds in order to develop an intuitive understanding of the kinds of images in the database.

There were 9 total conditions [(1/4, 1/2, or 1 leaf diameter) \times (full, texture removed, or texture only)]. Each condition of the experiment consisted of 620 trials in a single block, half of which were *same* and the rest *different*, and subjects were informed of this. On every trial, two image patches were displayed on the screen. During the first 20 trials, a large portion of the original image was also displayed next to the patches. These trials served to demonstrate the nature of the task to the subjects. In the subsequent 600 trials, only the two image patches were displayed. All data analyses were performed on these 600 trials. On each trial, subjects viewed the patches for as long as they desired and pushed a

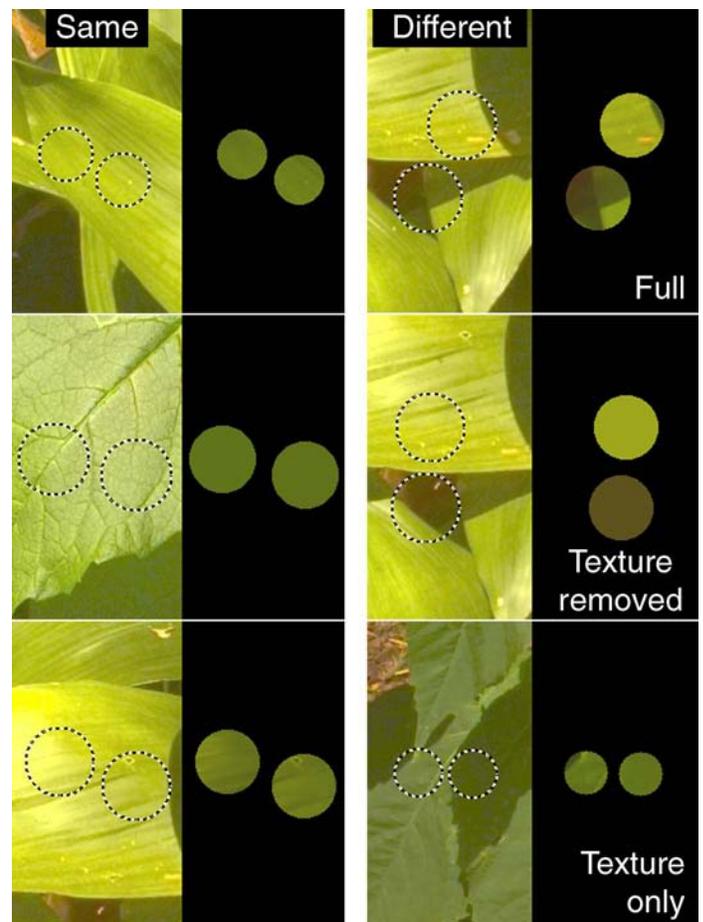


Figure 4. Six stimulus examples are illustrated. The patch pair is either *same* (left) or *different* (right). Both examples are taken from the *full* (row 1), *texture removed* (row 2), or *texture only* (row 3) condition with 1/4 leaf diameter between the patch centers. The patch pairs (on the right) were displayed on all 620 trials of the block, but the original image context (on the left) was only provided during the first 20 trials.

button to categorize the patch pair as *same* or *different* (Figure 4). Each of the 9 conditions was run in a separate 620-trial block, on separate days, and feedback was not provided.

	Average color		Color difference between patches			All contrast and spatial structure (texture) information		
	$\bar{\alpha}$	$\bar{\beta}$	ΔI	$\Delta\alpha$	$\Delta\beta$	$I^*(x,y)$	$\alpha^*(x,y)$	$\beta^*(x,y)$
Full	x	x	x	x	x	x	x	x
Texture Removed	x	x	x	x	x			
Texture Only	x	x				x	x	x

Table 2. This table illustrates the three experimental conditions under which human performance was measured in the psychophysical task. An x is present in each row where the selected dimension of information is present. Note that all contrast and spatial structure (texture) information is defined as a function of planar image coordinates (x, y); and when this information is removed, the patches are uniform. All three information sources are linearly separable and all are required to represent the image patches naturally.

After all conditions were run without feedback, corrective feedback was introduced and the conditions were run again. Feedback was given immediately after the subject categorized each stimulus. The correct category label was displayed onscreen and a tone indicated whether the response was correct or incorrect. Every 620-trial block was repeated (consecutively) at least once. The consecutive blocks continued until at least 3 blocks were run and until categorization accuracy (in the 600 trial block) stopped increasing.

Subjects

Two undergraduate females (“ska” age 18 yrs. and “cfa” age 21 yrs.) from the University of Texas at Austin were paid subjects in this experiment; neither was familiar with the scientific aims of the experiment. Both had normal color vision (Ishihara, 2001) and acuity ($\geq 20/20$).

Approximate ideal-observer models

An ideal observer performs at the theoretical upper limit in a given task. To evaluate the efficiency of human performance, and to quantify the potential usefulness of different image properties in the patch grouping task, we determined the performance of various ideal classifiers. Each of these ideal classifiers is defined by a specific combination of image properties used when making *same-different* classification decisions. The natural scene statistics measured in the current study consisted of the twelve image properties defined above (\bar{l} , $\bar{\alpha}$, $\bar{\beta}$, Δl , $\Delta\alpha$, $\Delta\beta$, $\Delta\sigma_l$, $\Delta\sigma_\alpha$, $\Delta\sigma_\beta$, $\ln F_l$, $\ln F_\alpha$, $\ln F_\beta$), and hence we only consider ideal classifiers that use different combinations of these image properties. Specifically, we constructed classifiers for each unique combination of the property sets defined in Equations 2, 3, 4, and 5 (μ_2 , μ_3 , Δ_3 , δ_3 , and F_3). For example, the classifier labeled $\mu_3\Delta_3\delta_3F_3$ used all 12 image properties in Equations 2, 3, 4, and 5, and the classifier labeled $\Delta_3\delta_3$ used the 6 image properties in Equations 3 and 4.

In the *texture-removed* conditions, the stimuli contained only five of these properties ($\bar{\alpha}$, $\bar{\beta}$, Δl , $\Delta\alpha$, $\Delta\beta$) and hence an ideal classifier that uses these properties ($\mu_2\Delta_3$) is the appropriate benchmark for comparison with human performance. Theoretically it is impossible for humans to outperform this ideal classifier in the *texture-removed* conditions, so if humans reach optimal performance levels, then humans can solve the task with perfect efficiency.

In the *full* and *texture-only* conditions, there are a large number of additional stimulus properties that humans could use in performing the task, and hence humans could potentially perform better than an ideal classifier that is limited to pick from the twelve properties that we

considered. If humans outperform an ideal classifier using the twelve properties, then we know that humans are using stimulus properties not in the set. Such an outcome would indicate that ideal performance can be improved by including more stimulus properties, and hence that it would be worth considering other potentially relevant stimulus properties.

It is appropriate to describe ideal classifiers within a Bayesian framework. In our specific patch grouping task, the goal is to maximize accuracy. (See the Discussion for discussion of another performance goal.) Because the prior probabilities of the two categories are equal, the ideal decision rule is to compute the posterior probability $p(\omega = \textit{same}|\mathbf{s})$ and response “*same*” if the posterior probability exceeds 0.5, which is equivalent to computing the log likelihood ratio:

$$z(\mathbf{s}) = \log \frac{p(\mathbf{s}|\omega = \textit{same})}{p(\mathbf{s}|\omega = \textit{different})}, \quad (6)$$

and responding “*same*” if $z(\mathbf{s}) > 0$, where \mathbf{s} is the stimulus on the trial. The log likelihood ratio function $z(\mathbf{s})$ will also be referred to here as the optimal decision function.

To compare human and ideal classifier performance it is critical that ideal performance be determined precisely. This is not trivial because there is always a limited amount of natural scene statistics data. There are many possible ways to deal with limited amounts of data. Here we considered two standard techniques (with a modification described later): Exemplar classification (e.g., Nosofsky, 1984) and quadratic classification (e.g. Ashby, 1992; Duda, Hart, & Stork, 2001). An exemplar classifier works by local probability density estimation (e.g., Duda et al., 2001; Parzen, 1962), and is conceptually similar to the k -nearest neighbor technique. A quadratic classifier is based on the assumption that the underlying probability density functions are Gaussian. Both the exemplar and quadratic classifiers specify a decision function that can be estimated from the natural scene statistics data. Under certain conditions (discussed below) these classifiers are optimal.

Exemplar classifier

An exemplar classifier is based on the idea of neighborhood density estimation for a large set of exemplars (stimuli) from each category. For any arbitrary D -dimensional stimulus $\mathbf{s} = (s_1 \dots s_D)$, the classifier “blurs” the exemplars surrounding \mathbf{s} to estimate the log likelihood ratio (Equation 6) of the two categories at \mathbf{s} . Here, we adopted an exponential blurring kernel because it is commonly used and because preliminary simulations suggested that it would work well for our data. The exemplar classifier has a free parameter for each stimulus

dimension, which determines the amount of blur along that dimension. Thus, the decision function is defined by

$$g(\mathbf{s}; \mathbf{w}) = \log \frac{\sum_{i=1}^N \exp \left[- \left(\sum_{d=1}^D [w_d (s_d - x_{same,i,d})]^2 \right)^{1/2} \right]}{\sum_{i=1}^N \exp \left[- \left(\sum_{d=1}^D [w_d (s_d - x_{diff,i,d})]^2 \right)^{1/2} \right]}, \quad (7)$$

where $\mathbf{w} = (w_1 \dots w_D)$ is the set of blurring parameters, $\mathbf{x}_{same,i}$ is the i^{th} exemplar in the training set from the category *same*, $\mathbf{x}_{diff,i}$ is the i^{th} exemplar from the category *different*, and N is the number of exemplars in each training set.

The exemplar classifier assumes that a sample of data can be blurred to determine the relative population density at any stimulus coordinate. As the sample size approaches infinity, neighborhood sampling noise approaches zero and therefore a narrow blurring kernel will provide accurate estimates of the relative population density for any stimulus coordinate (the blurring kernel approaches a delta function as the elements of \mathbf{w} approach infinity). In this case, the exemplar classifier will approach the Bayesian ideal $z(\mathbf{s})$ in Equation 6. However, under realistic conditions with finite sample sizes, the classifier performs best with some intermediate kernel size.

The exemplar classifier is useful because it adapts well to the coarse shape of most distributions. However, because the classifier relies on local blurring it will suffer from the effects of neighborhood sampling noise (the estimated distribution will have extra bumps and dips due to the finite sample size). This noise can be diminished by allowing flexibility in the size and shape of the blurring kernel, but it cannot be eliminated. The blurring kernel may also distort the true distributions by smoothing out regions where the distribution changes rapidly.

Determining the best performance of this classifier involves searching the space of possible blurring parameters. This can be very difficult in high dimensional spaces because of the effects of sampling noise, the shape of the blurring kernel, and the shape of true distribution functions. Below we describe an enhancement to the standard exemplar classifier that allows efficient estimation of the blurring parameters.

Quadratic classifier

If the stimulus density functions for the two categories [$p(\mathbf{s}|\omega = same)$ and $p(\mathbf{s}|\omega = diff)$] are Gaussian, then the optimal decision function is

$$q(\mathbf{s}; \mathbf{A}, \mathbf{b}, c) = \mathbf{sA}^T + \mathbf{bs}^T + c, \quad (8)$$

where the upper-triangular matrix \mathbf{A} , vector \mathbf{b} , and scalar c are free parameters that define the shape of a quadric surface defined over \mathbf{s} (Duda et al., 2001; Fisher, 1936). For example, if there are three stimulus dimensions, then

$$q(\mathbf{s}; \mathbf{A}, \mathbf{b}, c) = (s_1, s_2, s_3) \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} + (b_1, b_2, b_3) \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} + c.$$

Although the quadratic classifier was originally created for scenarios in which the stimuli in each category are Gaussian distributed (Fisher, 1936), the classifier is often applied in cases where the underlying distributions are not Gaussian.

The strengths and weaknesses of the quadratic classifier are opposite those of the exemplar classifier. The quadratic classifier is mostly immune to the effects of neighborhood sampling noise because $q(\mathbf{s}, \mathbf{A}, \mathbf{b}, c)$ does not reference data in local neighborhoods. Instead it asserts a global (quadric) shape for the decision function, but this also means that it is less flexible than the exemplar classifier. The quadratic classifier will perform well if the underlying distributions are similar to Gaussian (e.g., generalized Gaussian distributions), but it can perform poorly for more complex distribution shapes.

If the distributions are Gaussian then there are closed form expressions for the maximum likelihood estimates of the quadratic parameters (e.g., see Ashby, 1992; Duda et al., 2001). However, if the distributions are not Gaussian then the best performance of this classifier involves searching the space of possible quadratic parameters, which can be difficult depending on the number training samples and true shapes of the underlying distributions. Below we describe an enhancement to the standard quadratic classifier that allows efficient estimation of its parameters.

Enhanced implementation of exemplar and quadratic classifiers

The most straightforward way of implementing the exemplar and quadratic models is to estimate their parameters with an optimization algorithm such as gradient descent. However, we have found that such algorithms do not work well, presumably for the various reasons described in the previous two subsections. We obtained more robust results by employing a strong constraint that holds for optimal decision functions. Specifically, it is easy

to show (see [Appendix A](#)) that for an ideal classifier, the posterior probability of category membership given the stimulus is equal to the posterior probability of category membership given the value of the decision function: $p(\omega = \textit{same} | \mathbf{s}) = p(\omega = \textit{same} | z(\mathbf{s}))$. [Appendix A](#) also shows that the probability of category membership given the value of the decision function $p(\omega = \textit{same} | z(\mathbf{s}))$ is a non-decreasing function of the value of the decision function $z(\mathbf{s})$. Furthermore, this constraint holds for any decision function that is monotonic with $z(\mathbf{s})$.

We exploit this constraint by requiring that the posterior probability of $\omega = \textit{same}$ be a monotonic function of the decision variable. Let $\hat{z}(\mathbf{s})$ represent an estimate of the optimal decision function $z(\mathbf{s})$. Here, the exemplar $g(\mathbf{s}, \mathbf{w})$ and quadratic $q(\mathbf{s}, \mathbf{A}, \mathbf{b}, c)$ decision functions can be regarded as estimates of $z(\mathbf{s})$. For any given training data set consisting of N samples, we compute $\hat{z}(\mathbf{s})$ for each sample and then sort these $\hat{z}(\mathbf{s})$ values into quantiles. By definition, the quantiles contain equal numbers of samples. The j^{th} quantile will then contain $n_{j,\textit{same}}$ samples that are actually in category *same* and $n_{j,\textit{diff}}$ samples that are actually in category *different*. This provides an estimate $\hat{p}(\omega = \textit{same} | j)$ of the posterior probability of category *same* for each quantile of the decision variable:

$$\hat{p}(\omega = \textit{same} | j) = \frac{n_{j,\textit{same}}}{n_{j,\textit{same}} + n_{j,\textit{diff}}}. \quad (9)$$

The quantity $\hat{p}(\omega = \textit{same} | j)$ is a noisy estimate of the true posterior probability because it is subject to the effects of sampling noise and systematic errors in the estimated decision function $\hat{z}(\mathbf{s})$. For example, the blue curve in [Figure 5](#) illustrates the kind of non-monotonic relationship expected due to sampling noise. The blue curve was obtained by applying the optimal quadratic decision function to 10,000 random samples from Gaussian stimulus distributions and then binning the decision values into 200 quantiles to obtain $\hat{p}(\omega = \textit{same} | j)$. This number of samples is representative of our training data sets. The thick black curve shows the actual posterior probabilities that would be obtained with infinite sample size. The sampling noise apparent in the blue curve can make it difficult to search the parameter space of the decision function. To reduce the effects of sampling noise we enforce the non-decreasing constraint by finding the best fitting monotonic function $f(\omega = \textit{same} | j)$ through $\hat{p}(\omega = \textit{same} | j)$ by using a monotonic regression algorithm. This is illustrated by the red curve in [Figure 5](#), which is much closer to black curve (the actual posterior probabilities) than to the blue curve.

In determining the optimal parameters of exemplar and quadratic classifiers we minimized the following measure on the fitted non-decreasing function:

$$\bar{H} = -\frac{1}{2N} \sum_{\omega \in \{\textit{same}, \textit{diff}\}} \sum_{i=1}^N \log_2 f(\omega | J[\hat{z}(\mathbf{s}_{\omega,i})]), \quad (10)$$

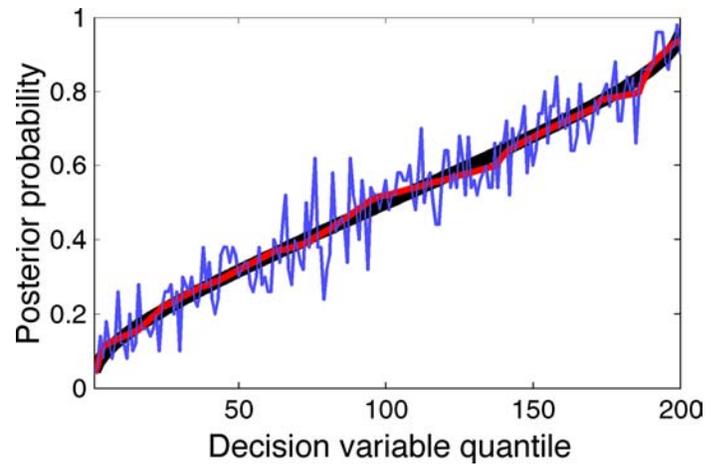


Figure 5. Illustration of the monotonic regression algorithm used to generate $f(\omega = \textit{same} | j)$ from $\hat{p}(\omega = \textit{same} | j)$, where j is the quantile number. Five thousand observations were sampled from two Gaussian categories with the distributions separated so that $d' = 1.0$. These samples were sorted into 200 quantiles (50 observations per quantile). The probability of category membership, $\hat{p}(\omega = \textit{same} | j)$, was computed for each quantile and is displayed as the blue curve. The monotonic regression algorithm was then applied to obtain $f(\omega = \textit{same} | j)$, which is displayed as the red curve. The black curve is the actual probability based on the known characteristics of the population. The monotonic regression reduced the mean squared error from the black curve by a factor of 20.

where $J[\hat{z}(\mathbf{s}_{\omega,i})]$ is the quantile of the value of the decision variable for stimulus $\mathbf{s}_{\omega,i}$, and N is the number of training stimuli in each category. The quantity \bar{H} is equivalent to the average value (over the quantiles) of the entropy of the posterior probability distribution of *same* and *different* in each bin (see [Appendix A](#)). The rationale for [Equation 10](#) is that entropy is a principled measure of the uncertainty associated with a probability distribution. At chance performance, $f(\omega = \textit{same} | j) = 0.5$, the entropy is 1 bit; and at perfect performance, $f(\omega = \textit{same} | j) = 1$ or $f(\omega = \textit{same} | j) = 0$, the entropy is 0 bits. Thus, \bar{H} decreases as the performance of the classifier improves. Compared to simply maximizing classification accuracy, we found that this measure reduces the tendency of the optimization algorithm to become trapped in local optima, while at the same time it yields final parameter values that generally maximize accuracy. Presumably, the average entropy measure performs better because it is less affected by sampling noise near the decision boundary.

Free parameters and cross-validation

Both the exemplar and quadratic classifiers have free parameters. If the number of stimulus properties is D (i.e., D is equal to the length of \mathbf{s}), then the exemplar classifier has D free parameters and the quadratic classifier

has $D(D + 3)/2 - 1$ free parameters. (The number of parameters for the quadratic classifier reflects the fact that, without loss of generality, the additive constant c can be set to zero and the vector \mathbf{b} can be scaled to a unit vector.)

To eliminate the effects of over-fitting we adopted a cross-validation approach (e.g., Efron & Tibshirani, 1993) to assess the performance of all 44 classifiers (22 quadratic and 22 exemplar). Since our database consists of 96 distinct images, each image served once as a test image while the remaining 95 images were used for parameter estimation.

Results

Image property distributions and ideal classification

As described earlier, a large number of approximate ideal classifiers were evaluated. Each of these classifiers was defined by the particular subset of the twelve stimulus dimensions (\bar{l} , $\bar{\alpha}$, $\bar{\beta}$, Δl , $\Delta\alpha$, $\Delta\beta$, $\Delta\sigma_l$, $\Delta\sigma_\alpha$, $\Delta\sigma_\beta$, F_l , F_α , F_β) used when making decisions in the patch grouping task. The performance of all the approximate ideal classifiers is summarized in the [Supplementary Material](#). In general, all of the quadratic classifiers achieved levels of performance that were approximately equal or slightly better than the corresponding exemplar classifiers. This suggests that the category distributions are amenable to quadratic classification and that the quadratic classifier's predictions are nearly optimal. In this subsection we present detailed results for four of these classifiers (when the distance between image patches is $1/4$ of the leaf diameter): Δ_3 , $\mathbf{s} = (\Delta l, \Delta\alpha, \Delta\beta)$; δ_3 , $\mathbf{s} = (\Delta\sigma_l, \Delta\sigma_\alpha, \Delta\sigma_\beta)$; F_3 , $\mathbf{s} = (F_l, F_\alpha, F_\beta)$; μ_3 , $\mathbf{s} = (\bar{l}, \bar{\alpha}, \bar{\beta})$. We also described results for several other classifiers and for other distances between the patches. Accuracy predictions for larger numbers of stimulus properties are given in the next subsection comparing ideal and human performance.

The first classifier, Δ_3 , only considers the differences in mean intensity and color between the two patches. The three bivariate scatter plots in [Figure 6a](#) show the pairwise distributions of intensity and color differences for patches from the same surface (green pixels) and from different surfaces (red pixels). The upper left plot shows the distributions for Δl , $\Delta\alpha$ (intensity vs. blue-yellow), the upper right plot for Δl , $\Delta\beta$ (intensity vs. red-green) and the lower plot $\Delta\alpha$, $\Delta\beta$ (blue-yellow vs. red-green). As expected, the differences between patches from the same surface tend to be more tightly clustered than those from different surfaces. This property of the image statistics can also be seen in the 1D marginal plots arrayed along the diagonal in [Figure 6a](#); the upper plot shows the marginal distributions for Δl (intensity), the middle plot for $\Delta\alpha$

(blue-yellow) and the lower plot for $\Delta\beta$ (red-green). The horizontal bars show the regions where one distribution dominates. We do not plot the full three dimensional distributions, but instead plot, in the lower left, the distributions of the posterior probabilities computed by the classifier for stimuli from the category *same* (green curve) and for stimuli from category *different* (red curve).

As shown in the lower left plot, the optimized quadratic classifier, using these three stimulus dimensions, has an accuracy of 78% correct in the patch grouping task (the exemplar classifier performs similarly with an accuracy of 76% correct). The other percentages in [Figure 6a](#) show the performance obtained when just two or one of the stimulus dimensions is used by the classifier. The contours in the scatter plots show the classifier decision boundaries. All three dimensions contribute similarly to overall performance, although the blue-yellow dimension is somewhat less useful.

[Figures 6b–6d](#) show the results for the other three classifiers. The second classifier, δ_3 , uses only the differences in intensity and color contrast between the two patches. This information is also useful, but the performance is poorer than for Δ_3 . The third classifier, F_3 , uses only the ratio of the intensity and color contrasts between the two patches. This classifier performs slightly less well than δ_3 , which uses contrast differences. Finally, the fourth classifier, μ_3 , uses only the average intensity and color of the two patches. We included this case because some of this information is available to the human observers, but these stimulus dimensions provide little useful information.

As might be expected, the performances of the classifiers decline systematically as the distance between the image patches increases to $1/2$ and 1 leaf diameter (see [Table 3](#)). However, the rank ordering of performance across the classifiers is preserved (see [Supplementary Material](#)), and the statistical distributions for *same* and *different* image patches are similar in shape to those in [Figure 6](#), although the overlap of the distributions increases with distance, primarily because of an increased spread in the *same* distributions.

The results in [Figure 6](#) imply that the information relevant to the patch grouping task is quite redundant across many of the stimulus dimensions. For example, as can be seen in [Figure 6a](#), Δl alone is sufficient for 74% correct performance, $\Delta\beta$ for 73% correct, $\Delta\alpha$ for 67%, and when all three dimensions are combined they yield 78% correct performance. However, if the three dimensions provided statistically independent information, then the accuracy would have reached 88% correct. Furthermore, when all of the contrast dimensions are combined with all of the intensity and color difference dimensions, performance increases to only 79% correct, and when all twelve stimulus dimensions are combined performance increases to only 80% correct (see [Table 3](#)).

Of the stimulus dimensions we analyzed, the intensity and color differences (Δl , $\Delta\alpha$, $\Delta\beta$) and contrast differences

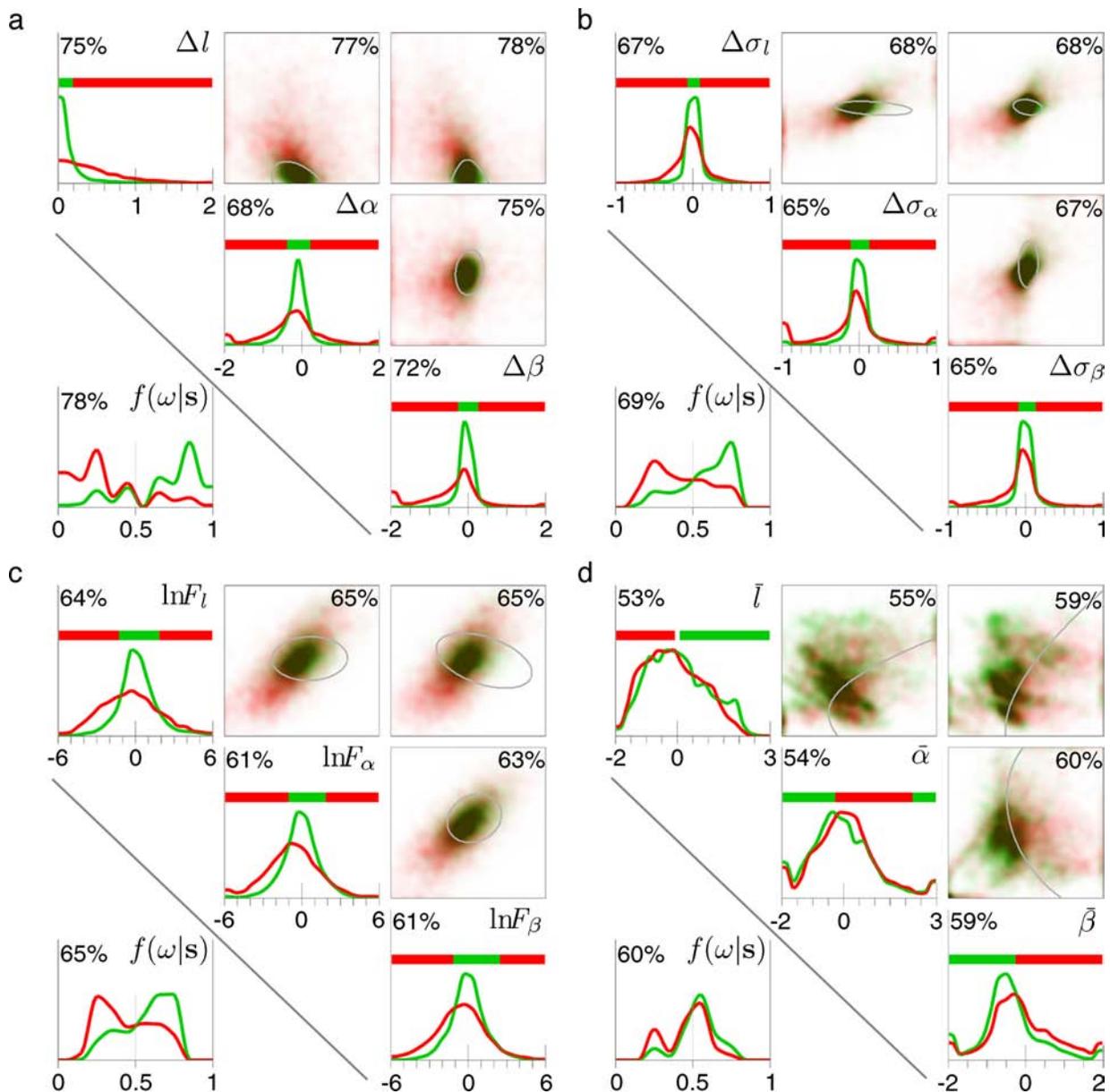


Figure 6. Image property distributions and quadratic classifier performance. **a.** Δ_3 Intensity and color differences. **b.** δ_3 Intensity and color contrast differences. **c.** F_3 Intensity and color contrast ratios. **d.** μ_3 Average intensity and color. Figures **a–d** show the distribution of each triple of image properties and show classification performance for the case where patches are separated by 1/4 leaf diameter. In each of these figures, the density of patch pairs for category *same* (green) and *different* (red) is shown. Along the diagonal, histograms for each individual image property are shown with the univariate classification rule (indicated by the red-green tape) and its associated accuracy. The horizontal axis represents the value of the image property and the vertical axis relative frequency. The upper-right triangle shows bivariate scatterplots of the data with quadratic decision boundaries depicted in gray. The horizontal and vertical axes of a bivariate plot are the same as the horizontal axes in the univariate plots below and to the left of the bivariate plot. (In the both the 1D and 2D plots the distributions are truncated, and the truncated mass is piled at the boundaries.) The full 3-dimensional distributions are not shown, but the lower-left corner of each figure shows histograms of the posterior probabilities computed by each trivariate classifier for *same* (green) and *different* stimuli. Also shown are the classification accuracy (percentages) obtained from cross-validation. The rank order of the classifiers from strongest to weakest is: Δ_3 , δ_3 , F_3 , μ_3 .

$(\Delta\sigma_l, \Delta\sigma_\alpha, \Delta\sigma_\beta)$ are the most useful for performance of the patch grouping task in close-up foliage, but even among these there is considerable redundancy so that good performance can be obtained with intensity differences alone (upper left plot in Figure 6a).

Figure 7 illustrates the results of the cross-validation analysis. In this analysis, the classifier parameters were determined separately for each test image to prevent overfitting. The univariate histograms in Figure 7 show the distribution of accuracy across the 96 test images for the

	Accuracy			Mean Entropy		
	1/4	1/2	1	1/4	1/2	1
$\mu_3\Delta_3\delta_3F_3$	0.80	0.75	0.70	0.68	0.76	0.85
$\mu_2\Delta_3\delta_3F_3$	0.79	0.75	0.70	0.68	0.77	0.84
$\mu_3\Delta_3\delta_3$	0.80	0.75	0.70	0.68	0.76	0.85
$\mu_3\Delta_3F_3$	0.80	0.75	0.70	0.70	0.76	0.85
Δ_3	0.78	0.74	0.67	0.72	0.81	0.91

Table 3. Some results from the cross-validation procedure are shown. Mean accuracy and mean entropy \bar{H} reflect the average performance of five quadratic classifiers, where patches are separated by 1/4, 1/2, and 1 leaf diameter.

five different classifiers in Table 3. As can be seen, performance varied widely across the individual images. The bivariate scatter plots compare the accuracies of the five classifiers. If one classifier is superior to another, then more of the points will fall on its side of the diagonal in the scatter plot. Figure 7 shows, in agreement with Table 3, that adding stimulus dimensions to Δ_3 produces small (but significant) improvements in performance.

Human classification performance

The symbols in Figure 8 show the classification accuracy of the two naïve subjects. Each panel of

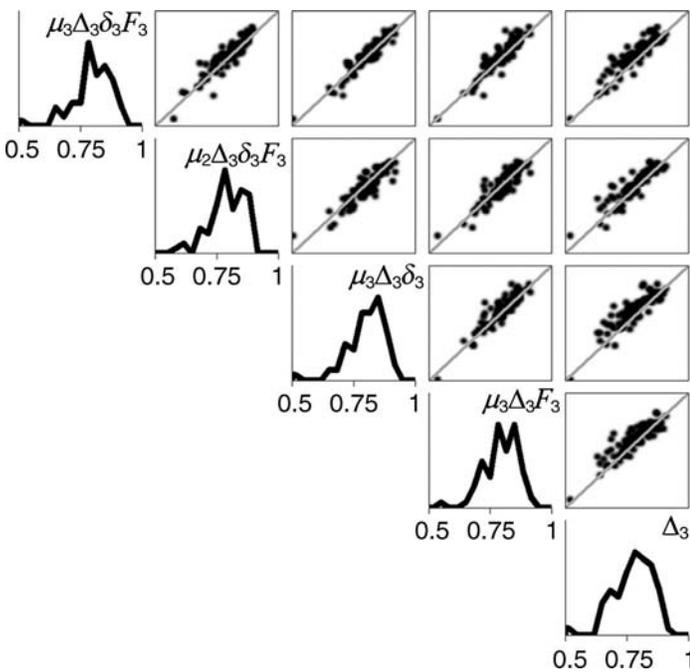


Figure 7. Accuracy of five quadratic classifiers for each of the 96 test images where patch pairs were separated by 1/4 leaf diameter. Accuracies were obtained with the cross-validation procedure (i.e., test images were not in the training set). The marginal distributions of these scores are shown along the diagonal and the pair-wise distributions (taken two classifiers at a time) are shown in the upper-right triangle.

Figure 8 plots accuracy as a function of the distance between the image patches for a different kind of image information. Recall that in the *full* condition the image patches were taken directly from the natural images (except for a scaling to bring them into the range of the monitor), in the *texture-removed* condition the image patches retained their mean color as well as their intensity and color differences, and in the *texture-only* condition the image patches retained everything in the *full* condition except for the intensity and color differences. The green symbols show the raw accuracy levels. The black symbols show the accuracy corrected for bias, based on standard signal detection analysis (Green & Swets, 1966). This corrected accuracy is given by

$$p_{C_{\max}} = \Phi \left[\frac{\Phi^{-1}(p_s) + \Phi^{-1}(p_d)}{2} \right], \quad (11)$$

where p_s and p_d are subject's probability correct for *same* and *different* stimuli, $\Phi(\cdot)$ is the standard normal integral function, and $\Phi^{-1}(\cdot)$ its inverse. As can be seen, the performance of the two observers was highly correlated ($r = 0.84$), accuracy declined as a function of the separation between the image patches, was similar for the *full* and *texture-removed* conditions, and poorer for the *texture-only* condition.

The gray curves show quadratic classifier performance. For the *full* conditions, the model classifier used intensity and color differences, contrast differences, contrast ratios and means ($\mu_2\Delta_3\delta_3F_3$); for the *texture-removed* conditions, the model classifier used only the intensity and color differences and means ($\mu_2\Delta_3$); and for the *texture-only* conditions, the model classifier used only the contrast differences, contrast ratios and the means ($\mu_2\delta_3F_3$). Note that only for the *texture-removed* conditions is the model classifier guaranteed to be close to the true ideal classifier. This cannot be guaranteed in the *full* and *texture-only* conditions because the model classifiers were restricted to use the magnitudes of patch contrast while humans were shown the exact spatial pattern of the patches.

Figures 8a and 8c show human and classifier accuracy for the initial phase of the experiment where feedback was not given. These results show that the accuracy of the

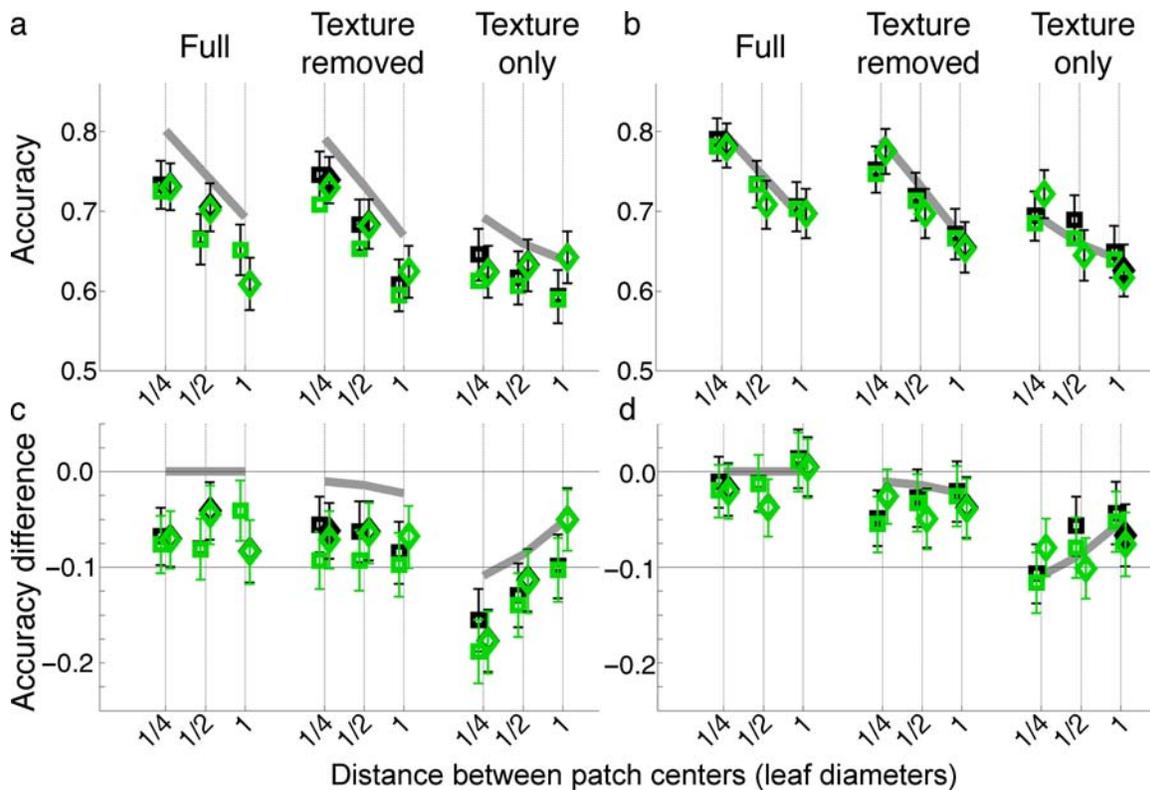


Figure 8. Performance in the patch grouping task for two human observers (symbols) and an ideal classifier (gray curves). **a.** Human accuracy without feedback compared to ideal. **b.** Human accuracy with feedback (after practice) compared to ideal. The green symbols represent raw human accuracy and the black symbols represent accuracy after correction for bias. **c.** and **d.** Same data as in **a** and **b** plotted as the difference in accuracy from the ideal classifier that uses all cues. (Error bars indicate 90% confidence intervals.)

human observers parallels but is slightly below that of the classifiers. For the *texture-removed* conditions the results show that human efficiency is relatively high. For the other conditions, little can be concluded about efficiency because the model classifiers do not incorporate all the potential sources of information available to the human observers.

Figures 8b and 8d show the human and classifier accuracy for the second phase of the experiment where feedback was given on each trial. Again, the performance of the two observers was highly correlated ($r = 0.92$), but the performance of both subjects improved ($p < 0.01$ by a bootstrap test). Here human performance has improved and is closer to the model classifiers. This result suggests that although the subjects entered the experiment with good knowledge of the local statistics of foliage images, they were able to adapt to the task with training.

One important result is that human performance in the *full* and *texture-only* conditions does not exceed the performance of model classifiers that were restricted from using any texture and spatial pattern information except the magnitudes of patch contrast. This suggests that perhaps high order texture and spatial pattern information

is not used by human observers in this specific task. Another potential way of detecting whether humans are using such information is to compare their performance on small and large image patches. The smallest quartile of image patches in the experiment had diameters ranging from 15 to 23 minutes of arc and the largest quartile had diameters ranging from 43 to 195 minutes of arc. Presumably the larger patches contain more high order information. If humans use this information in the task, then we might expect them to perform better than model classifiers on larger patches. Figure 9 shows the raw accuracy scores for stimuli containing the smallest quartile of image patches (blue symbols) and the largest quartile (red symbols). There is little or no improvement in accuracy as a function of patch size in the *full* and *texture-removed* conditions, but in the *texture-only* conditions, human and model classifier accuracy increased by similar amounts. The curves show the performance levels of the model classifiers for large and small patch sizes (these are the same classifiers used for Figure 8). As the figure shows, these classifiers, which are restricted from using any texture properties except the magnitudes of patch contrast, can account for the change in accuracy (as a function of patch size) shown by humans. These results

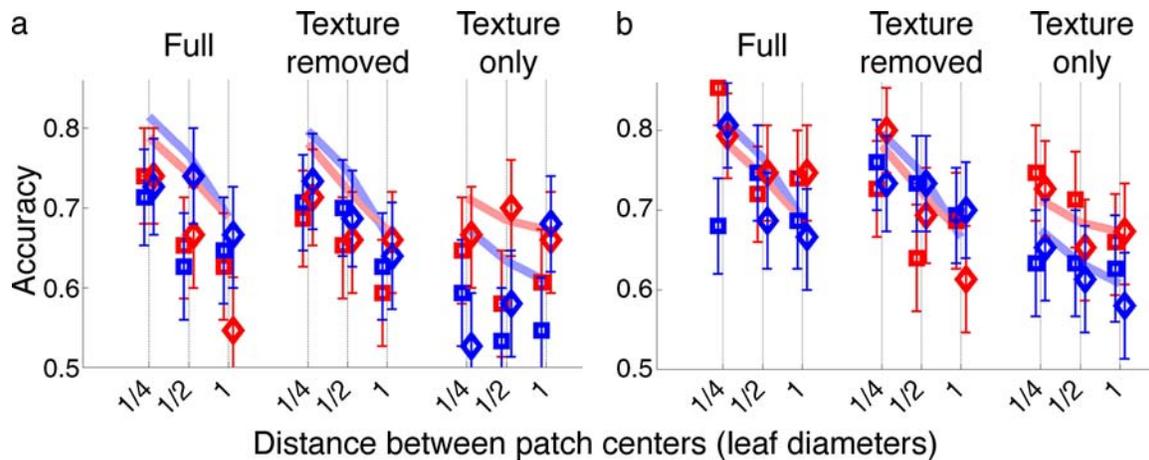


Figure 9. Performance in the patch grouping task for small (blue symbols) and large (red symbols) image patches. **a.** Human accuracy without feedback compared to ideal. **b.** Human accuracy with feedback compared to ideal. The curves show the classifier performance and the symbols show human accuracy. (The error bars represent 90% confidence intervals.)

add further evidence that humans do not make use of texture and spatial pattern information in our patch grouping task.

Discussion

The aims of this study were (a) to measure the statistics of some of the local image properties in natural foliage scenes that might support grouping and segregation into regions corresponding to leaf surfaces, (b) to determine the decision rules that an optimal visual system (with knowledge of these statistics) should use to perform a simple patch grouping task, and (c) to compare optimal and human performance in this patch grouping task.

Patch-pair statistics

We measured the statistics for twelve different intensive and contrast measures for patches within and across leaf boundaries. Some of the marginal and pair-wise distributions for these properties are shown in Figure 6. We found that differences in mean intensity/color (Figure 6a) and differences in intensity/color contrast (Figures 6b and 6c) were smaller within a leaf surface than across a leaf surface boundary, and thus both intensive and contrast differences could provide useful information for region grouping and segregation. The distributions of mean intensity and color of the patch pairs were similar within surfaces and across surfaces boundaries (Figure 6d), and thus these properties are not as useful.

Optimal classifier performance

To determine quantitatively how useful the patch-pair statistics might be for region grouping and segregation, we

attempted to determine the performance of optimal classifiers (ideal observers) in a simple patch grouping task where on each trial, the classifier is presented with two patches and must decide whether they belong to the *same* or *different* surfaces. We evaluated two kinds of classifier: A quadratic classifier that is guaranteed to be optimal when the underlying distributions are Gaussian, and an exemplar classifier that (with sufficient data) can outperform the quadratic classifier if the underlying distributions are not Gaussian. Both classifiers performed similarly (see [Supplementary Materials](#)), suggesting that both classifiers approached optimal performance for the local image properties that we analyzed. We found that differences in mean intensity/color (ΔI , $\Delta \alpha$, $\Delta \beta$) were the most effective properties for solving the task, the contrast differences ($\Delta \sigma_I$, $\Delta \sigma_\alpha$, $\Delta \sigma_\beta$) were the next most effective, the contrast ratios (F_I , F_α , F_β) were the next most effective, and the overall average intensity/color (\bar{I} , $\bar{\alpha}$, $\bar{\beta}$) were the least effective. Furthermore, all 11 classifiers that included properties in addition to the differences in mean intensity/color performed only slightly better than the classifier that included only the differences in mean intensity/color. For all of these classifiers, accuracy increased from approximately 70% correct when patches were separated by 1 leaf diameter to approximately 80% correct when patches were separated by $1/4$ leaf diameter (chance = 50%). The fact that performance is better at smaller distances suggests that there may be advantages to using region-growing mechanisms when performing region grouping and segregation.

An obvious question is whether the choice of color space affects classifier performance. To investigate this question we determined classifier performance for three other color spaces:

1. the opponent color space used by Johnson, Kingdom, and Baker (2005),

2. the CIE $L^*u^*v^*$ and
3. the CIE $L^*a^*b^*$ spaces (e.g., see Wyszecki & Stiles, 1982).

Perhaps not surprisingly, we found that performances were similar (within a couple of percent correct) for all the color spaces. Nonetheless, the $l\alpha\beta$ space consistently gave the best performance.

Monotonic regression and the calculation of category likelihoods

In the methods section we introduced a monotonic regression procedure that is useful for estimating classifier parameters for our patch grouping task. The procedure involves quantiling the classifier's decision function values and fitting a monotonic function to the posterior probabilities of category membership across quantiles. We found that this procedure reduces the likelihood of becoming trapped in local optima by reducing the random sampling effects of a finite training set. Here we describe another benefit of monotonic regression in designing and implementing optimal classifiers.

Most natural tasks involve making inferences about the environment from sensory stimuli. Sensory stimuli generally contain many different sources of information that are relevant to performing a natural task, and hence optimal performance often requires combining many sources of information. In an optimal classifier, each source of information typically takes the form of posterior probabilities over possible states of the environment given the observed values of selected stimulus properties. The proposed monotonic regression procedure can increase the precision of posterior probability estimates by reducing sampling noise and systematic error. This is illustrated in Figures 5 and 10, which shows the results of three simulations. As described earlier, Figure 5 illustrates the result of using the monotonic regression procedure to estimate the parameters of the quadratic classifier when

the underlying distributions are Gaussian. The red curve shows the estimated posterior probability function, which gives the probability of category *same* as a function of the value of the optimized quadratic decision variable. The black curve shows the actual posterior probability function based on the underlying Gaussian distributions. The close agreement between the red and black curves is not surprising given that the quadratic decision function is the optimal decision function when the underlying distributions are Gaussian—the quadratic decision function is the logarithm of the likelihood ratio of the underlying Gaussian distributions (see Equations 6 and 8).

Figure 10 illustrates the result of using the monotonic regression procedure to estimate the parameters of the quadratic classifier when the underlying distributions are not Gaussian. Figure 10a shows the results when the distributions are generalized Gaussian (e.g., see Box & Tiao, 1992) with an exponent of 1.0 (heavier tails than Gaussian), and Figure 10b shows the results with an exponent of 4.0 (more rapid falloff than Gaussian). The red curves show the estimated posterior probability function and the black curves show the actual posterior probability function based on the underlying non-Gaussian distributions. Importantly, the gray curves show the posterior probabilities that would be calculated directly from the optimized quadratic classifier parameters under the standard assumption that the underlying distributions are Gaussian. Interestingly, the quadratic decision function still produces optimal accuracy in the classification task for these non-Gaussian distributions (after parameter optimization). However, the quadratic decision function does not correspond to the log of the likelihood ratio of the non-Gaussian distributions, and hence (as Figure 10 illustrates) it cannot be directly used to obtain accurate estimates of the posterior probabilities for different observed values of the decision variable. In these cases, when the monotonic regression procedure is used in conjunction with the quadratic decision function, a much more accurate posterior probability function is obtained (as shown by the red curves). Of course, for some underlying distribution

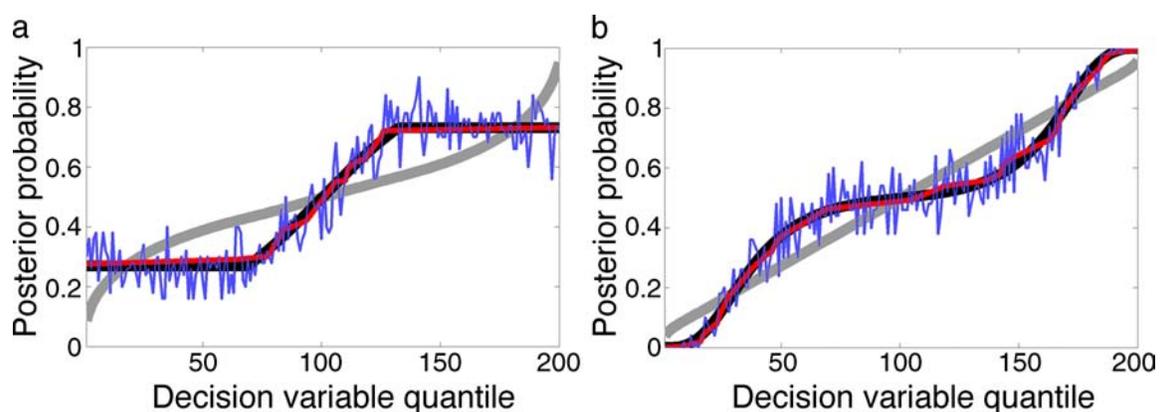


Figure 10. Simulation of monotonic regression procedure for non-Gaussian distributions, where accuracy corresponds to 70% correct ($d' = 1.0$). **a.** Generalized Gaussian with exponent 1.0. **b.** Generalized Gaussian with exponent of 4.0. See Figure 5 for further explanation.

shapes, the quadratic decision function will be suboptimal and another kernel function can be substituted (e.g., see Shawne-Taylor & Cristianini, 2004). Whatever families of decision functions are considered, coupling them with the monotonic regression procedure is likely to yield more robust and accurate estimates of the posterior probability functions.

Human versus ideal performance

To compare human and ideal performance, human performance was measured in the patch grouping task in two phases. The first phase was administered without corrective feedback and the second phase included trial-by-trial feedback. In the first phase, human performance paralleled but fell slightly below that of the best model classifiers (Figure 8a). In the second phase, human and model classifier performances were even more congruent (Figure 8b). There are two important implications of these results. First, the middle panels of Figures 8a and 8b show that humans efficiently use the mean intensity/color difference information in performing the patching grouping task. Second, the left panels in Figures 8a and 8b show that human performance does not improve substantially when given all the spatial-pattern/texture information available in the image patches. Similarly, the right panels show that in *texture-only* conditions humans do not perform better than a model classifier that uses only simple contrast information. These results (especially the good match with ideal after practice) suggest either that like the model classifiers, the human observers only use simple intensive and contrast differences or, more likely, that the higher order spatial pattern information (e.g., surface markings, shading and shadow patterns, etc.) is not of much value in this particular task.

Human and model classifier performance did not exceed 80% correct in the patch grouping task. While this is a reasonable performance level, it falls well below what is possible with full context, as is obvious from inspecting the larger images (e.g., see Figures 2, 4, and 11). Indeed, in a preliminary experiment, we found that when the stimuli are displayed with their original context (similar to the appearance of Figure 4 but with a larger context area) and when humans were free to change their viewing distance from the display, they were able correctly classify 620 out of 620 stimuli. What is responsible for the modest performance with the image patches in the current experiment? Some intuition can be obtained by examining those images for which the model classifiers performed poorly versus those for which they performed better. Row 1 in Figure 11 shows portions of images for which the model classifiers are less than 65% correct at the $\frac{1}{4}$ diameter distance. These images contain strong shadows, specular highlights and/or different colored regions within a surface. Row 2 shows portions of images for which classifier performance was approximately 80%

correct. These images may contain surface markings, minor shading and lighting features and/or similar intensity and color across the surface boundaries. Finally, row 3 shows portions of images where the classifiers performed at or above 90% correct. In these images the leaves are fairly uniform and the intensity and color differences across the surface boundaries are relatively strong.

These examples suggest that one of the major reasons humans perform well with larger image regions is that they are able to determine which image contours are due to shadows, shading and other lighting effects, which are due to surface reflectance changes, and which are due to surface boundaries. A small image patch may not contain sufficient information to determine which kind of contours it contains (McDermott, 2004), and thus accurate region grouping may require detecting and interpreting extended contours in addition to growing regions on the basis of local color and texture similarity. Thus, an obvious next step is to statistically analyze the properties of natural images that support contour detection and integration and that allow determination of whether a contour is due to a change in surface reflectance, lighting, or a surface boundary.

Relation to previous studies

The present study is closely related to a previous study by Fine et al. (2003). They measured the joint distribution of pixel color differences, $p(\Delta I, \Delta \alpha, \Delta \beta)$, in the Ruderman et al. color space, as a function of the spatial distance between pixels within natural images, as well as across completely different images. They assumed that neighboring pixels in the same image are random samples from the same physical surface, pixels from different images are random samples from different surfaces, and that pixels at some separation within an image are random samples from a mixture of both same and different surfaces. These assumptions allowed them to determine for any pair of pixels (at any spatial separation) the posterior probability that they were drawn from the same or different surfaces. They then compared the posterior probabilities generated by this statistical model with judgments of human subjects in a task where the subjects were asked to judge whether or not each pixel in a randomly sampled patch of natural image belonged to the same surface as the pixel at the center of the patch. They found a modest correlation between the human judgments and the predictions of the statistical model, but not nearly as strong a correlation as that observed in the current study (see texture-removed conditions in Figures 8 and 9). There are at least two potential reasons for this difference. One is that the image patches they presented to subjects included more contextual information. Thus, the subjects' judgments may have been strongly influenced by natural image statistics that were not represented in their model. Another potential

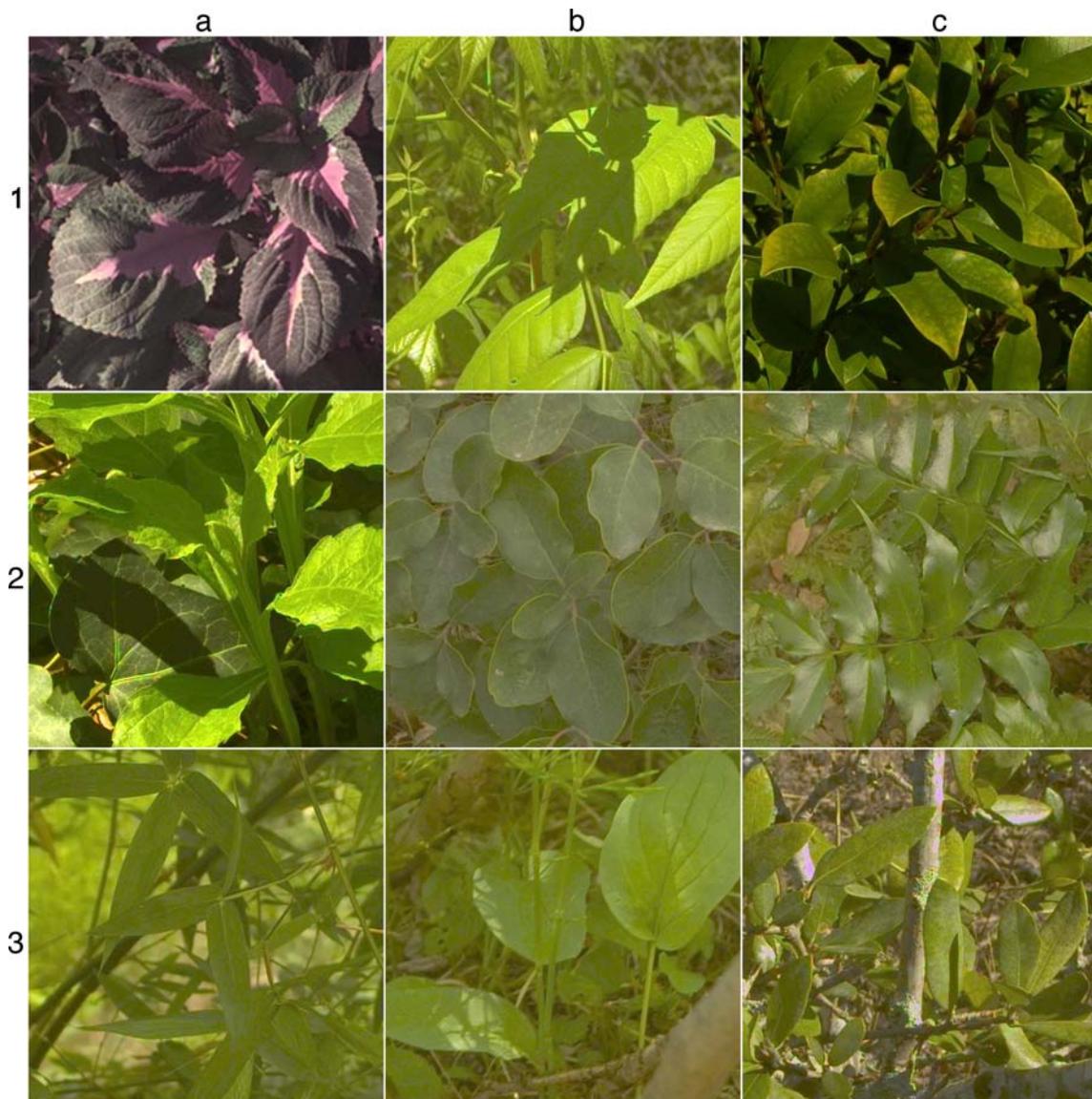


Figure 11. Example portions of natural foliage images. **Row 1.** Portions of images for which the model observers perform relatively poorly. The leaves contain strong shadows, specular highlights, and/or differently colored regions within a surface. Image 1a clearly stood out as the most difficult image, as accuracy was just above 50%. **Row 2.** Portions of images for which the model observers performed moderately well. The leaves tend to contain surface markings, minor shading features and specular highlights, or have similar intensity and color to adjacent surfaces. **Row 3.** Portions of images for which the model observers performed relatively well. These leaves tend to be fairly homogenous and distinct from neighboring regions in terms of intensity.

reason is that their model is based on the assumption that the joint distribution of color differences does not depend on the spatial distance between pixels that fall within the same surface. We found that this assumption is violated in close-up foliage images. In fact, the strong dependence of the joint distributions on distance within surfaces is the reason that the models in the current study predict performance to decline as function of the distance between image patches (see [Figures 8 and 9](#)).

In previous studies from our lab, we measured contour statistics in natural images, derived ideal observers, and then compared human and ideal performance in a simple

contour-occlusion task, where the observer must decide when two contour elements passing under an occluding region belong to the same or different contours (Geisler & Perry, 2009; Geisler, Perry, & Ing, 2008). As in the present study, we found that human performance closely parallels that of the ideal observer. Somewhat differently, we found no evidence of practice effects (improvements in performance) with trial-by-trial feedback. We concluded in the previous study that the visual systems of naïve human subjects contain rather complete knowledge of the pair-wise statistical properties of natural contours. The present study suggests that although naïve subjects have

good implicit knowledge of the statistics of intensity and color differences in image patches from close-up foliage, they can increase that knowledge with practice. This difference between the two tasks may reflect differences in the importance of the image properties for scene interpretation or in the generality of the statistics. It may be, for example, that the statistics of contour shape are more useful and more consistent across various scales and classes of natural environment than are intensity and color differences.

Appendix A

Monotonic constraint

In a binary classification task the decision rule that maximizes accuracy is to compute the posterior probability of each category given the observed data and then pick the category with the highest probability. In the present case, this corresponds to computing $p(\omega = \textit{same} | \mathbf{s})$ and then responding “*same*” if $p(\omega = \textit{same} | \mathbf{s}) > 0.5$. A simple application of Bayes’ formula shows that for equal prior probabilities of the two categories this rule is equivalent to computing the log likelihood ratio

$$z(\mathbf{s}) = \log \frac{p(\mathbf{s} | \omega = \textit{same})}{p(\mathbf{s} | \omega = \textit{diff})} = \log \frac{p(\omega = \textit{same} | \mathbf{s})}{p(\omega = \textit{diff} | \mathbf{s})}, \quad (\text{A1})$$

and responding “*same*” if $z(\mathbf{s}) > 0$. It follows that $p(\omega = \textit{same} | z(\mathbf{s})) = p(\omega = \textit{same} | \mathbf{s})$, because all points \mathbf{s}_1 in the set defined by $z(\mathbf{s}_1) = z_1$ have the same posterior probability $p(\omega = \textit{same} | \mathbf{s}_1)$. Furthermore, it follows from [Equation A1](#) that if $z(\mathbf{s}_2) > z(\mathbf{s}_1)$ then $p(\omega = \textit{same} | \mathbf{s}_2) > p(\omega = \textit{same} | \mathbf{s}_1)$ and hence $p(\omega = \textit{same} | z(\mathbf{s}_2)) > p(\omega = \textit{same} | z(\mathbf{s}_1))$.

Average entropy

From text [Equation 10](#) we have that

$$\begin{aligned} \bar{H} = & -\frac{1}{2N} \sum_{i=1}^N \log_2 f(\omega = \textit{same} | J[\hat{z}(\mathbf{s}_{\textit{same},i})]) \\ & -\frac{1}{2N} \sum_{i=1}^N \log_2 f(\omega = \textit{diff} | J[\hat{z}(\mathbf{s}_{\textit{diff},i})]). \end{aligned} \quad (\text{A2})$$

If m is the number of quantiles, then the number of samples in a quantile is $n = 2N/m$, and hence

$$n = n_{j,\textit{same}} + n_{j,\textit{diff}}. \quad (\text{A3})$$

Thus, we can rewrite [Equation A2](#) as

$$\bar{H} = -\frac{1}{nm} \sum_{j=1}^m n_{j,\textit{same}} \log_2 f(\omega = \textit{same} | j) + n_{j,\textit{diff}} \log_2 f(\omega = \textit{diff} | j). \quad (\text{A4})$$

But, $f(\omega = \textit{same} | j) = \frac{n_{j,\textit{same}}}{n}$ and $f(\omega = \textit{diff} | j) = \frac{n_{j,\textit{diff}}}{n}$, and therefore

$$\begin{aligned} \bar{H} = & -\frac{1}{m} \sum_{j=1}^m f(\omega = \textit{same} | j) \log_2 f(\omega = \textit{same} | j) \\ & + f(\omega = \textit{diff} | j) \log_2 f(\omega = \textit{diff} | j), \end{aligned} \quad (\text{A5})$$

which is the average across bins of the entropy in each bin.

Acknowledgments

Supported by NIH EY11747.

Commercial relationships: none.

Corresponding author: Wilson S. Geisler.

Email: geisler@psy.utexas.edu.

Address: Mezes Hall 330, Austin, TX, USA.

References

- Ashby, F. G. (1992). Multivariate probability distributions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 1–34). Mahwah, NJ: Erlbaum.
- Balboa, R. M., & Grzywacz, N. M. (2000). Occlusions and their relationship with the distribution of contrasts in natural images. *Vision Research*, 40, 2661–2669. [[PubMed](#)]
- Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.
- Brunswik, E., & Kamiya, J. (1953). Ecological cue-validity of “proximity” and other Gestalt factors. *American Journal of Psychology*, 66, 20–32. [[PubMed](#)]
- DiCarlo, J. M., & Wandell, B. A. (2000). Illuminant estimation: Beyond the bases; Paper presented at IS&T/SID Eighth Color Imaging Conference; Scottsdale, AZ (pp. 91–96).
- Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. Second Edition, New York: John Wiley and Sons.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. London: Chapman & Hall.

- Elder, J., & Goldberg, R. (2002). Ecological statistics of Gestalt laws for the perceptual categorization of contours. *Journal of Vision*, 2(4):5, 324–353, <http://journalofvision.org/2/4/5/>, doi:10.1167/2.4.5. [PubMed] [Article]
- Fine, I., MacLeod, D. I. A., & Boynton, G. M. (2003). Surface segmentation based on the luminance and color statistics of natural scenes. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 20, 1283–1291. [PubMed]
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.
- Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure/ground cues are valid for natural images. *Journal of Vision*, 7(8):2, 1–9, <http://journalofvision.org/7/8/2/>, doi:10.1167/7.8.2. [PubMed] [Article]
- Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59, 10.1–10.26. [PubMed]
- Geisler, W. S., & Diehl, R. L. (2003). Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, 27, 379–402. [PubMed] [Article]
- Geisler, W. S., & Perry, J. S. (2009). Contour statistics in natural images: Grouping across occlusions. *Visual Neuroscience*, 26, 109–121. [PubMed] [Article]
- Geisler, W. S., Perry, J. S., & Ing, A. D. (2008). Natural systems analysis. In B. Rogowitz & T. Pappas (Eds.), *Human vision and electronic imaging, SPIE Proceedings* (vol. 6806). Bellingham, WA: SPIE.
- Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41, 711–724. [PubMed]
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Ishihara, S. (2001). *Ishihara's tests for colour deficiency, Concise Ed.* Tokyo: Kanehara Trading Inc.
- Johnson, A. P., Kingdom, F. A. A., & Baker, C. L. (2005). Spatiochromatic statistics of natural scenes: First- and second-order information and their correlation structure. *Journal of the Optical Society of America A*, 22, 2050–2059.
- Kersten, D., Mamassian, P., & Yuille, A. L. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, 55, 271–304. [PubMed]
- Knill, D. C., & Richards, W. (Eds.) (1996). *Perception as Bayesian inference*. Cambridge, MA: Cambridge Univ. Press.
- Konishi, S., Yuille, A. L., Coughlan, J. M., & Zhu, S. C. (2003). Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 57–74.
- Krinov, E. (1947). Nation Research Council of Canada, Ottawa.
- Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 530–549. [PubMed]
- McDermott, J. (2004). Psychophysics with junctions in real images. *Perception*, 33, 1101–1127.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 104–114. [PubMed]
- Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematics Statistics*, 33, 1065–1076.
- Reinagel, P. (2001). How do visual neurons respond in the real world? *Current Opinion Neurobiology*, 11, 437–442. [PubMed]
- Ruderman, D. L., Cronin, T. W., & Chiao, C. (1998). Statistics of cone responses to natural images: Implications for visual coding. *Journal of the Optical Society of America A*, 15, 2036–2045.
- Shawne-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge, UK: Cambridge University Press.
- Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion Neurobiology*, 13, 144–149. [PubMed]
- Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Reviews of Neuroscience*, 24, 1193–1215. [PubMed]
- Stockman, A., MacLeod, D. I. A., & Johnson, N. E. (1993). Spectral sensitivities of the human cones. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 10, 2491–2521. [PubMed]
- Torralba, A. (2009). How many pixels make an image. *Visual Neuroscience*, 26, 123–131. [PubMed]
- Wyszecki, G., & Stiles, W. S. (1982). *Color science: Concepts and methods, quantitative data and formulae*. New York: John Wiley and sons.