

Spatial biases and computational constraints on the encoding of complex local image structure

Ryan R. L. Taylor

ARC Centre of Excellence in Vision Science and Centre for
Visual Sciences, Research School of Biological Sciences,
Australian National University,
Canberra, Australia



Ted Maddess

ARC Centre of Excellence in Vision Science and Centre for
Visual Sciences, Research School of Biological Sciences,
Australian National University,
Canberra, Australia



Yoshinori Nagai

Center for Information Science, Kokushikan University,
Tokyo, Japan



The decomposition of visual scenes into elements described by orientation and spatial frequency is well documented in the early cortical visual system. How such 2nd-order elements are sewn together to create perceptual objects such as corners and intersections remains relatively unexplored. The current study combines information theory with structured deterministic patterns to gain insight into how complex (*higher-order*) image features are encoded. To more fully probe these mechanisms, many subjects ($N = 24$) and stimuli were employed. The detection of complex image structure was studied under conditions of learning and attentive versus preattentive visual scrutiny. Strong correlations ($R^2 > 0.8$, $P < 0.0001$) were found between a particular family of spatially biased measures of image information and human sensitivity to a large range of visual structures. The results point to computational and spatial limitations of such encoding. Of the extremely large set of complex spatial interactions that are possible, the small subset perceivable by humans were found to be dominated by those occurring along sets of one or more narrow parallel lines. Within such spatial domains, the number of pieces of visual information (pixel values) that may be simultaneously considered is limited to a maximum of 10 points. Learning and processes involved in attentive scrutiny do little if anything to increase the dimensionality of this system.

Keywords: spatial statistics, isotricon textures, entropy, spatial vision, attention, learning, visual cortex, computational modeling

Citation: Taylor, R. R. L., Maddess, T., & Nagai, Y. (2008). Spatial biases and computational constraints on the encoding of complex local image structure. *Journal of Vision*, 8(7):19, 1–13, <http://journalofvision.org/8/7/19/>, doi:10.1167/8.7.19.

Introduction

Visual structure is described by sets of relationships between multiple points within an image. Recently it has been demonstrated that in order to accurately reconstruct important local image features such as corners or intersections, sets of spatial interactions between four or more points (higher-order spatial correlations) need to be considered (Franz & Schölkopf, 2006; Victor, Conte, Purpura, & Katz, 1995). In studying how these image features are visually encoded, one must go beyond traditional typically 2nd-order tools such as gratings (Simoncelli & Olshausen, 2001). Unfortunately higher-order spatial correlations are hard to quantify and experimentally control. These difficulties have placed strong constraints on studies into the visual processing of complex spatial information. To a large degree, the present work overcomes such problems by combining information theoretic measures of

visual information with an extension of the pioneering methods of Julesz (Julesz, Gilbert, & Victor, 1978; Maddess, Nagai, Victor, & Taylor, 2007).

The current approach allowed us to probe two important issues. The first is whether local form analysis is restricted to relatively low order spatial interactions. That is, to test the dimensionality of the system. The second is an attempt to determine how local signals might be spatially combined so as to capture higher-order image structure. The roles of learning and attentive versus preattentive processing on these fundamental elements are also investigated.

To study how higher-order image structure is encoded, one needs a sufficiently rich collection of stimuli. To this end, a large battery of isotricon texture stimuli (Maddess, Nagai, James, & Ankiewicz, 2004; Maddess et al., 2007) were employed. The battery consisted of 21 texture types or ensembles, each containing individual *ternary* (composed of 3 contrast values, $-1, 0, 1$) textures generated by

simple arithmetic rules applied recursively to initially random patterns. Single examples from each ensemble are given in Figure 1. Importantly, even for small collections of these patterns (>10) the average third-order (and lower) correlation functions (3CFs) of each ensemble are not significantly different from zero (Maddess et al., 2007). This is also the case for uniformly distributed noise patterns (pixel values assigned randomly to -1 , 0 , 1 with equal probability). This means the isotrison textures are completely isotropic when considering measures that are third-order and below (see Methods). To discriminate such ensembles from each other and from noise, one must therefore learn ensemble-specific higher-order features (Victor, 1994; Victor et al., 1995). The majority of neurons in primate V1 have been shown to be sensitive to structure defined at fourth order and above when stimuli that are able to quantify this have been used (Purpura, Victor, & Katz, 1994). In the present case, the average probability of correctly differentiating each of the ensembles used here from noise textures provides a basis for relating particular measures of image information to processes in the visual system. Of central concern is how and to what extent a large range of complex spatial structures (defined by 4th and higher spatial correlations) are encoded. As such, the present investigation lies in the domain of form perception rather than texture processing, where for example the spatial integration of simple element properties such as orientation (Beck, Sutter, & Ivry, 1987; Field, Hayes, & Hess, 1993; Landy & Bergen,

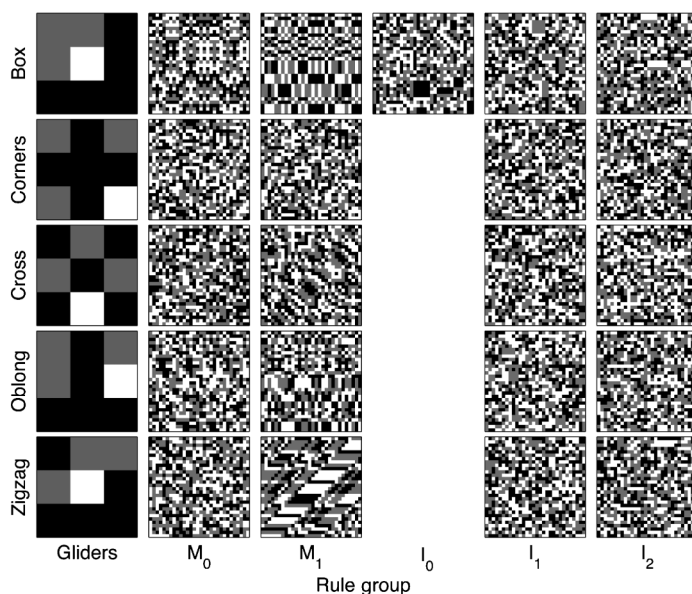


Figure 1. Single texture examples from each of the 21 ensembles examined. Textures within each ensemble are generated by a simple set of arithmetic rules that determine how pixel values falling under the inputs of a glider (gray squares in glider column) are combined to generate an output (white square) as the given glider and rule is recursively applied to uniform ternary noise.

1991) or periodicity (von der Heydt, Peterhans, & Dürsteler, 1992) are more of interest.

The number of unique combinations of pixels (i.e., spatial structures) that may be observed within a randomly selected 3×3 pixel region, or clique, of a 3-level noise pattern is $3^9 = 19683$, while for isotrison ensembles there are between 3 and 81 times fewer observable cliques (Maddess et al., 2007). In this regard, these textures are more natural than noise having similarly low dimensionality to natural images (Chandler & Field, 2007). The present ensembles also share a third-order property (evenly distributed bispectral “energy”) that is highly characteristic of regions of natural images that are fixated upon by human-observers (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000). Given a much lower number of possible combinations of pixel values, it might be expected that all ensembles should appear more structured than noise patterns. Despite such statistical structure, psychophysics reveals that many ensembles are not differentiable from random patterns (Maddess & Nagai, 2001; Maddess et al., 2007). This suggests that some higher-order features are more readily detected than others.

These biases are perhaps ecologically driven and may expose some of the underlying features of complex form processing. By varying the shape of the pixel sampling domains (e.g., to some shape other than the 3×3 example above) used to calculate ensemble or image information (entropy) and comparing those measures with human data for each ensemble, it is possible to learn something of both the range and spatial biases of local form processing. Those results suggest that processing of complex higher-order relationships may be biased toward interactions occurring along one to three parallel one dimensional horizontal or vertical spatial domains. The results also suggest that processes involved in learning and attentive visual scrutiny do little to increase the order of local form analysis which seems to reach capacity at about 10th-order pixel interactions. In sum, the present study may provide significant insights into the local computational and spatial limits of visual form processing.

Methods

Higher-order stimuli: Texture ensembles

Figure 1 presents single examples of each of the 21 texture types tested. Briefly, such textures are generated by a recursive process whereby a *glider* (Figure 1, left column) is passed over each pixel of an initially *ternary* (composed of three luminance values, -1 , 0 , 1) evenly distributed random pattern (Maddess & Nagai, 2001; Maddess et al., 2004, 2007; Victor & Conte, 1991). As each randomly assigned pixel falls under the output pixel

of the glider, its value may be changed depending on the values of the pixels falling under the input pixels and one of a set of rules governing how inputs are combined to determine an output value. The particular rules insure the higher order properties of the textures (Maddess et al., 2007). The present texture types, also referred to as ensembles, are generated by 5 gliders (Figure 1) and 5 isotrison rules M_0 , M_1 , I_0 , I_1 , and I_2 (Maddess et al., 2007). Texture ensembles generated by the same glider and different rules share some important relationships (Maddess et al., 2004).

The present test battery includes almost all ensembles generated by 3 input gliders (Maddess et al., 2004). In order to reduce the length of the experiment, some ensembles that cannot be discriminated from random were excluded. Nonetheless, many ensembles that appear random were included. This exhaustive set was chosen because it contains a very large variety of structure. For example, all possible ternary 3×3 combinations of pixel values (cliques) ($N = 19683$) occur within this set. An additional reason for employing the present ensembles is that they have been studied previously and therefore facilitate direct comparison between the present work and other studies.

To underline the defining statistical property of isotrison textures, ensembles and evenly distributed noise patterns cannot be differentiated by their average third or lower order (power spectrum and average brightness) spatial structure. This statistical property defines the current ensembles as *isotrison* and is described by the 3rd order correlation function $C_{3,f(h_1,v_1,h_2,v_2)}$ (3CF) for an image $I(x, y)$ comprised of N pixels each with an area of 1

$$C_{3,f(h_1,v_1,h_2,v_2)} = \frac{1}{N} \sum \sum I(x,y) I(x+h_1,y+v_1) \times I(x+h_2,y+v_2) \Delta x \Delta y. \quad (1)$$

The 3CF gives the average correlation (products) between all possible triplets of pixels, where the product is between a pixel (x, y) and two others at horizontally (h) and vertically (v) shifted locations. The 3CF is thus the third order analogue of the second order correlation function (2CF), i.e., the Fourier transform of which is the power spectrum of $I(x, y)$. Explained in detail below, the mean 3CF of isotrison ensembles is everywhere zero (as with evenly distributed noise patterns) (Maddess et al., 2004, 2007). To reiterate, this equality implies that only fourth and higher-order information can be used to identify a pattern as belonging to a particular ensemble or noise.

Many of the textures used here look quite structured and their isotrison nature indicates that the structure is due to correlations at fourth order and above (Maddess et al., 2004, 2007). We have stated that the patterns show no significant mean orientation bias, i.e., they are isotropic, for lower order measures like the 2CF or 3CF. This is true even for as few as 10 to 12 examples, as shown in recent

studies of their convergence (Maddess et al., 2007). One feature of the textures used in our psychophysical tests that is clearly anisotropic is the checks from which the patterns are composed. Clearly the checks have horizontal and vertical orientation biases. For the entropy calculations described below, the tessellated elements, e.g., checks, are irrelevant since only the values of the element colorings, i.e., whether the elements (checks) are black, white, or gray (contrasts -1 , 0 or 1), enter the calculations. It is important to note that for the psychophysical stimuli the bias from the texture elements is at spatial frequencies higher than the inverse of the distance between the tessellated elements, whatever their shape. This can be seen in Figure 2.

The isotrison texture type that seems to generate the greatest suspicion of being anisotropic at lower orders are the binary even Box textures, so a demonstration that their apparent anisotropic character is not due to the checks, or to structure at lower correlations would be appropriate. Each row of Figure 2 presents results for 16 different Box textures, 16 being the minimum number seen by any subject in this study. The textures come in 5 variants, the left-most column shows an example of each. The top row describes the original textures presented at the size and check number used in the psychophysical experiments. The next row down describes properties of blurred Box textures, and the subsequent rows describe textures where the original checks are replaced by other elements: smaller circles, squares, and diamonds, but where the colorings are still those of Box textures. The second column, entitled “elements,” shows the mean of the power spectra of the tessellated element of each texture variant, except for the blurred textures where the spectrum is that of the large 2D Gaussian operator that produced the blur. The next column, labeled “textures,” contains the mean power spectra of each set of textures. The next column, labeled “central” is the central portion of each spectrum. All the spectra shown in these columns have been raised to the power of 0.2 to emphasize any high frequency content relative to low.

Several points are clear from these examples (Figure 2). First, information about the coloring of the checks is at frequencies below the inverse of the check size or 1.85 cpd ($= 1/0.54$). The spectrum above that frequency is determined by the shape of the tessellated elements. Some structure is visible in the region below 1.85 cpd but this is different for each average spectrum, i.e., there is no consistent structure. Also none of the coefficients is significant, the t -values computed for each of the 2560 (512×5) coefficients having a mean of 0.71 ± 0.16 SD. Our brains do not have the spatial frequency resolution of these spectra. To provide a more realistic model, power spectra were calculated using channels covering the same central frequencies for channels having bandwidths of 0.8 cpd. The coefficients are presented as t -statistics for each region (Maddess et al., 2007) as shown in the rightmost column of Figure 2. The mean t -statistic for

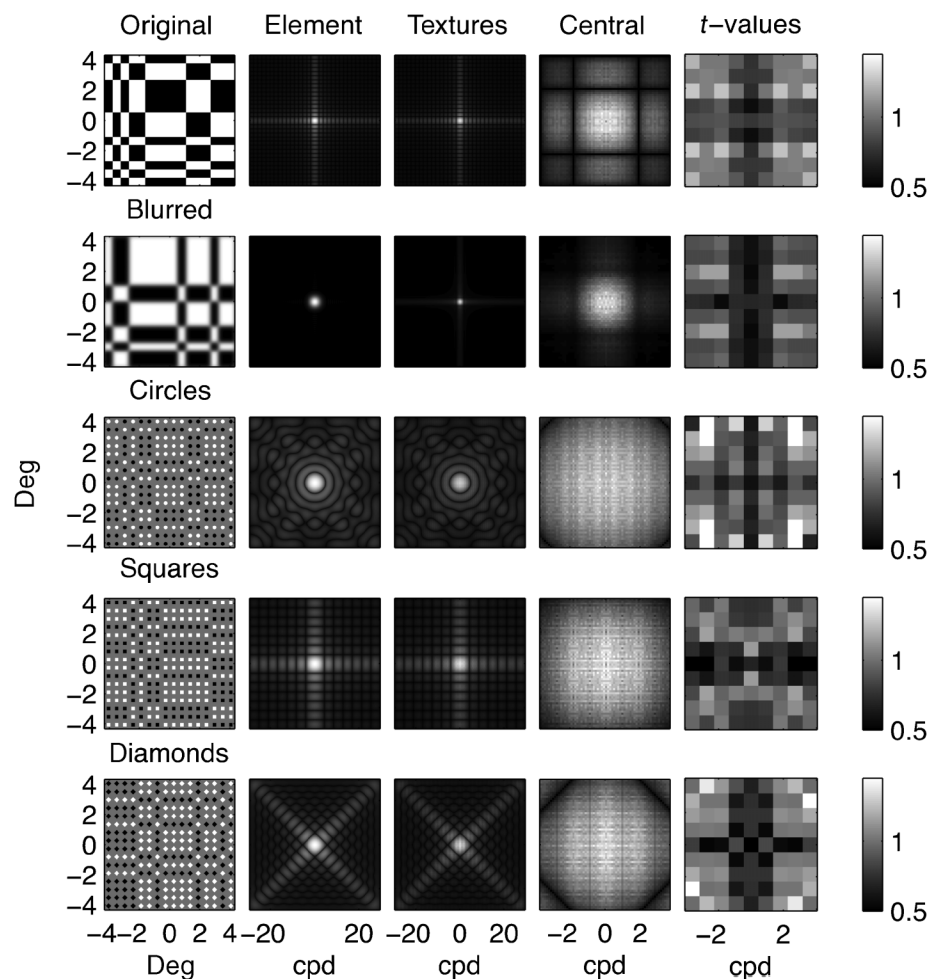


Figure 2. Textures are isotropic (similar in all orientations) at third-order and below. Each row presents average results from 16 binary Even Box textures. Five variants (rows, single examples in first column) are investigated to illustrate that the apparent anisotropy of these highly structured patterns is not due to their component checks or to lower-order correlations. The average power spectrum of the tessellated element of each variant is given in the second column (“elements”). The average tessellated power spectrum for each set of variants is given in the third column (“textures”). The central portion of these spectrums are given in the fourth column (“central”). All the spectra have been raised to the power of 0.2 to emphasize high relative to low frequency content. The fifth column presents t -statistics for power spectra coefficients calculated for the central frequencies from channels with bandwidths of 0.8 cpd.

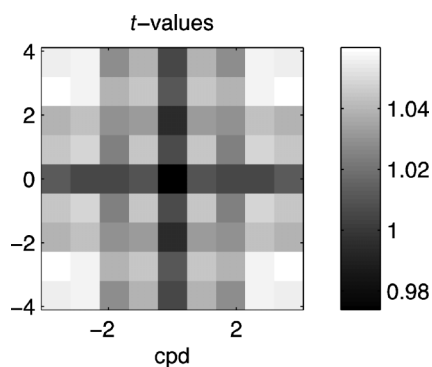


Figure 3. Isotropy holds for averages based on very small sets of examples. Average t -values were calculated for 300 sets of spectra coefficients averaged over 3 textures. At no orientations are the coefficients significant. Therefore, even for small sets of textures there is no orientation bias.

horizontal and vertical coefficients for the 80 texture examples was 0.71 ± 0.15 SD with a maximum of value 1.15; in other words, no coefficient was significantly different from 0. Note also that the larger t -values tend to originate from where the power spectrum is smallest. Even if the coefficients were significant they do not show a preponderance of horizontal or vertical elements, i.e., they are isotropic for this second order measure.

While the textures are isotropic for measures computed at third order and below, and non-significant for an average of 16, or somewhat fewer examples (Maddess et al., 2007), one might conjecture that significance might occur for averages of fewer textures. To explore this we computed the t -values for sets of 3 spectra (Figure 3). We examined 300 such sets of t -values, the mean being presented in Figure 3. Here the t -values are somewhat larger (but recall that for $N = 3$ larger t -values are required to reach significance), the mean spectrum remaining

isotropic. The same was true for amplitude spectra rather than the power spectra. Hence, even for small collections there is no mean orientation bias. One could possibly look for significant horizontal or vertical components in pairs of different frequency bands in particular examples; however, the action of comparing power_1 AND power_2 is formally fourth order. Hence, while by definition there may be an orientation bias at fourth order or above there is none below that, even for these most unisotropic looking textures.

It is also worth noting that discrimination of the textures from noise patterns has been studied at a range of check sizes (Maddess & Nagai, 2001; Victor & Conte, 1989). At the smallest size studied, a degree of optical blurring would have occurred (Maddess & Nagai, 2001; Victor & Conte, 1989) making the spectrum of the stimuli look like that of the blurred stimuli presented here. Note that even when raised to the power of 0.2 there is barely any information apparent above the inverse of the check size (e.g., Figure 2, row 2, column 4). In those experiments, performance was invariant with check size indicating that discrimination was mainly due to the coloring of the checks not any higher frequency components (when they were visible). Taken together those results, and the similarity of the psychophysical results presented here to the entropy measures (which know nothing of the checks), indicate that the present psychophysical tests were also mainly determined by the colorings and not the check properties. As shown in Figures 2 and 3, there is no orientation bias that can be picked up in the coloring of the isotrigran patterns except for measures that operate at fourth order and above.

Materials and stimulus presentation

Achromatic stimuli were presented on a model CCID 7551 monitor (Barco, Kortrijk, Belgium) at a refresh rate of 101 Hz, with a resolution of 512×420 square pixels and a mean luminance of $45 \text{ cd} \cdot \text{m}^{-2}$. Testing was conducted in a darkened room in which ambient light was provided by the display monitor. Using a chin rest, subjects were required to fixate on a small dot in the center of the screen viewed binocularly at a distance of 60 cm. All software was written in Matlab (Matlab; The MathWorks, Natick, MA).

For investigations of preattentive visual processing, stimuli were presented for 297 ms, less than the time required for focused attentive scrutiny (Maddess & Nagai, 2001). For the attentive processing tests, each stimuli was presented for 30 seconds. For each presentation, the contrast of the test stimulus was increased from 0 to 1 and subsequently decreased onset and offset phases following a Blackman function (Maddess & Nagai, 2001). Each ternary pattern, either noise or a member of an ensemble, was a region comprised of 16×16 black, white or gray checks centered in the middle of the screen.

Each check covered approximately $0.54^\circ \times 0.54^\circ$ of the retina. In keeping with the approach of previous studies (Maddess et al., 2007; Victor & Mast, 1991) patterns were presented within an evenly distributed ternary random background of 64×52 checks.

Psychophysics

We tested 24 subjects, 9 of whom were highly familiar with the isotrigran texture discrimination tasks, and 15 for whom the tests were novel. Each subject completed 16 two alternative forced choice trials (2AFCs) for each of the 21 ensembles. Each trial required subjects to decide whether a presented pattern was either an evenly distributed random pattern or belonged to the ensemble being tested. Before each block of 16 2AFC trials, subjects were given 20 examples from the test ensemble and 20 evenly distributed noise patterns to study. All these texture examples were 33 pixels square presented on a gray background. To facilitate learning during 2AFC trials, feedback on performance was provided by a tone that sounded for incorrect choices (entered by either a left or right mouse click). A total of 9408 responses were recorded from three groups of subjects: preattentive inexperienced (15 subjects), preattentive experienced (9 subjects) and attentive experienced (7 subjects).

Spatially biased measures of higher-order image information

A benefit of the present texture ensembles is that all unique configurations of pixel values ($-1, 0, 1$) observed within some specified spatial arrangement (a “Sampler”), such as a 1D strip of 9 pixels (see Figure 4i), are equally likely. This fact reduces measures of image information to a statement about the number of unique combinations of pixels that may be observed under a sampler placed on a given ensemble. Equi-probability was confirmed by sampling several million examples for each of many samplers, although this is the theoretically expected outcome for these textures (Gilbert, 1980). For example, if each of all observed N pixel configurations (*words*) is equi-probable (normal for these textures) ($p(x_i) = 1/N$), the equation for information entropy is simplified to

$$H(X) = - \sum_{i=1}^N p(x_i) \log_2 p(x_i) = -1 \times \log_2 \left(\frac{1}{N} \right) = \log_2(N) (\text{bits}). \quad (2)$$

As a check the more general entropy (Equation 2, left) was also computed in each case and was found to be highly consistent with the simpler equi-probable version. Information theoretic measures also allow quantitative

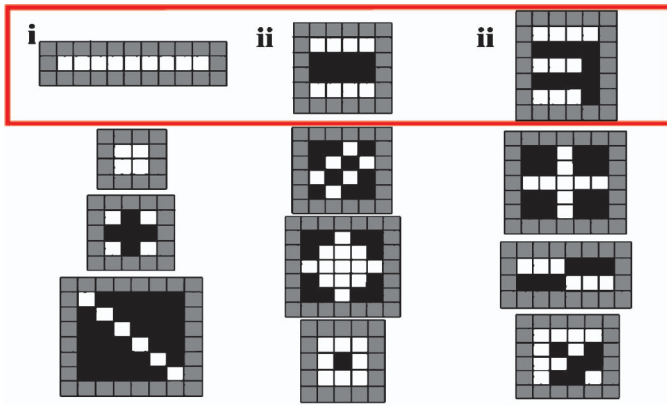


Figure 4. Examples of spatial configurations (“samplers”) employed to investigate image complexity. Only the bright pixels of the sampler domain were selected to form a word, which was then used as an index into the histogram of observed words. More than 800 samplers were investigated, including randomly generated samplers and those created by algebraic rules that insure significant differences between individual samplers (e.g., orthogonal, independent). Samplers that best matched the preattentive inexperienced, experienced, and attentive data are boxed (i, ii, and iii, respectively). Other examples are “blobs” (row 3, column 2), the so-called even/odd domains (row 2, column 1), and one of 36 6×6 Welch/Hadamard functions (row 4, column 4).

comparisons between ensembles, to this end conditional entropy was also computed ($H(X|Y)$) (Cover & Thomas, 1991). The shapes of samplers, including their number of input pixels (Figure 4), were varied and the measured ensemble information compared with obtained psychometric functions. The sampling process is illustrated in Figure 5 where a sampler is placed on an example from the M_0 oblong ensemble (see Figure 1). The pixel combination, or word, is then recorded and the process repeated by shifting the sampler along to adjacent pixels. This process continues until the sampler has covered the entire texture, another example from the same ensemble is then analyzed in the same fashion. Multiple examples are sampled in this manner, ending when all words within an ensemble have been identified. The final result is the number of unique words (N) occurring in a given ensemble and consequently the entropy of the ensemble (Equation 2).

The present analysis seeks samplers that detect structure (lower ensemble entropy) to an extent similar to human observers. That is, samplers that yield high entropy for those ensembles that humans have difficulty discriminating, and low entropy when humans find the discrimination easy, are deemed to yield similar performance to humans (i.e., lower complexity = more perceptible). Thus, we tended to compare $1/\text{entropy}$ with discrimination probability. Those samplers that yielded measures of ensemble complexity similar to human performance may inform us both about the dimensionality of the visual system’s

analysis and how pixel information might be spatially pooled.

If there is a particular spatial bias in the processing of higher-order interactions, then ensembles that share a large number of words should appear perceptually similar. An obvious counter balance to this similarity is the number of words not shared, for example, ensemble X may contain every word in Y , although such words may constitute only a small fraction of the total observable in X (i.e., X has higher complexity). Both these factors are captured by a simple dimensionless measure ($S(X, Y)$) of the similarity between pairs of ensembles. The measure gives the ratio of the words shared between X and Y and the total number of unique words found in both X and Y :

$$S(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}. \quad (3)$$

That is, $S(X, Y)$ is the cardinality (number of words) of the intersection of X and Y ($|X \cap Y|$) divided by the cardinality ($| \cup |$) of the union X and Y ($|X \cup Y|$). Where no words are shared $|X \cap Y| = 0$, and therefore $S(X, Y) = 0$. Conversely, where both ensembles share all their words $S(X, Y) = 1$. Therefore, in terms of shared words, a value close to 1 denotes high similarity between ensembles, while a value close to zero denotes little similarity. If a particular sampler perfectly captures the range over which local interactions are taken then $S(X, Y) = 1$, and it should be impossible to discriminate between examples of X and Y .

As the number of testable samplers grows exponentially with the number of input pixels they contain (e.g., for 3-level textures there are 3^{25} possible words in a 5×5 clique) only a subset of the total number were tested. For sample domain shapes and sizes ranging from lines of 4 pixels to 6^2 pixel matrices, this group included vertical,

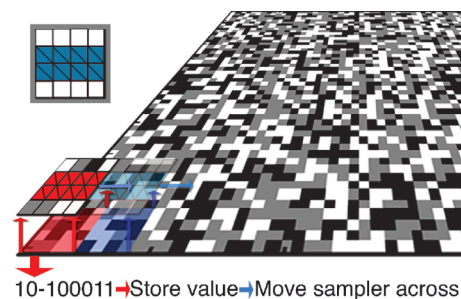


Figure 5. Illustration of the sampling process for a single texture example. A sampler (inset left, input pixels are white) is passed across a 150×150 element texture from the M_0 Box ensemble. The values of the pixels falling underneath the samplers inputs are recorded before the sampler is moved to the next location on the texture. The process continues for multiple examples stopping when each word has been counted several thousand times. Given that each word is equal probable, the final result is simply the number of unique words counted for each ensemble.

diagonal, and horizontal bar patterns, checker board patterns, all 36 matrices orthonormal of 6×6 Walsh/Hadamard functions (e.g., Figure 4, bottom right). The four input sampler in the second row and first column corresponds to an often explored type of easily perceived higher-order structure (e.g., Beason-Held et al., 1998; Victor & Conte, 2005). As this set reflects a diverse range of different configurations, at least one should be somewhat similar to the true shape of the human higher-order spatial bias if it exists, if so further “tuning” of the search is possible by focusing on variants of the successful samplers.

Results

Psychophysics: Learning and preattentive vs. attentive processing

As reported in other isotrigon/isodipole studies, performance was found to vary widely between ensembles and less so between subjects (Maddess & Nagai, 2001; Maddess et al., 2007; Victor & Mast, 1991). Figures 6A and 6B compare performance between the different subject groups for each of the 21 ensembles tested. Figure 6C gives an example of how performance for a typical subject (f21a) varied between preattentive and attentive conditions. Notable in all data sets is that some gliders (e.g., Box) and rules (e.g., M_1) produce more discriminable textures.

Neither experience (Figures 6A and 6B gray lines) nor conscious scrutiny (black line, Figure 6B) led to greatly improved performance relative to naïve preattentive viewing (black line, Figure 6). This may not be surprising where discrimination performance is either at a maximum (e.g., M_1 Box textures) or there is no discernable structure (e.g., I_1 Oblong) and therefore there is nothing to learn. It is interesting that for some ensembles performance might decrease with conscious scrutiny (Figure 6B; I_2 ZigZag, I_0 Box), although this was not significant when multiple comparisons were taken into account. Also notable are ensembles where it appears that average performance is below chance (e.g., M_0 Cross).

Ensemble entropy and conditional entropy

Of the large number of samplers investigated the most “human-like” are given in Figure 7 (A to C). Here $1/\text{normalized entropy}$ (obtained by dividing each entropy value by the highest entropy value) (black) is equated with discrimination accuracy (psychometric function, PF, gray). The results suggest a clear preference for interactions occurring along some number of single pixel wide strips. In the case of inexperienced subjects (Figure 7A, gray), the best matching sampler (Figure 7A, black) is a single horizontal strip of 9 pixels ($R^2 = 0.81$, $P < 0.001$) (Figure 7A right). Experienced (Figure 7B, gray) preattentive data are best matched by entropy measures based on pairs of short horizontal strips ($R^2 = 0.81$, $P < 0.001$). The attentive scrutiny data (Figure 7C, gray) are most correlated ($R^2 = 0.78$, $P < 0.001$) with measures based on

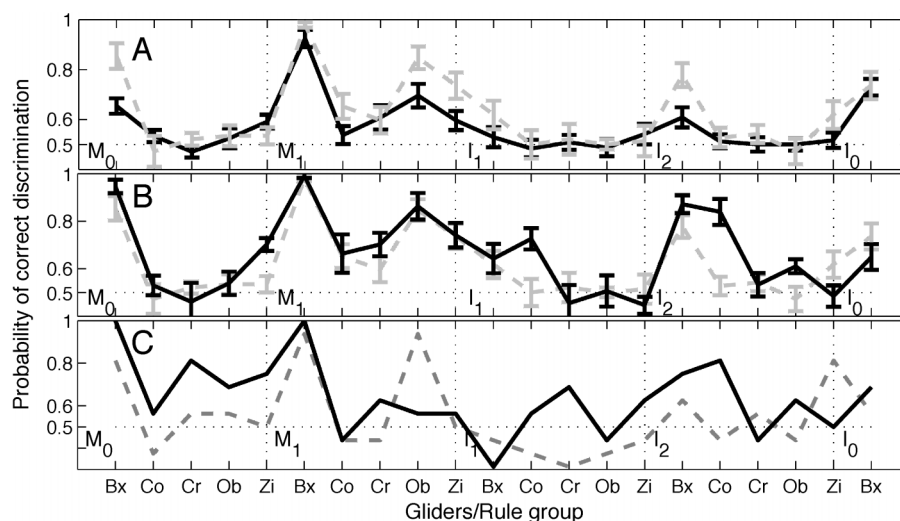


Figure 6. Psychometric functions (PFs). Group averaged performances (A and B) and a single subject example (C). (A) PFs for the preattentive discrimination task, in which subjects had to decide whether a briefly presented (200 ms) texture was uniform noise or belonged to a given ensemble. Average ($\pm$ SE) inexperienced subject (black line) performance is plotted with the average experienced (gray line) PF. (B) Compares average performance ($\pm$ SE) of experienced subjects in the preattentive experiment (gray line) to the same subjects given 30 seconds to view ensembles (black line). (C) Preattentive (gray) vs. attentive (black) performance of a typical experienced subject (f21a).

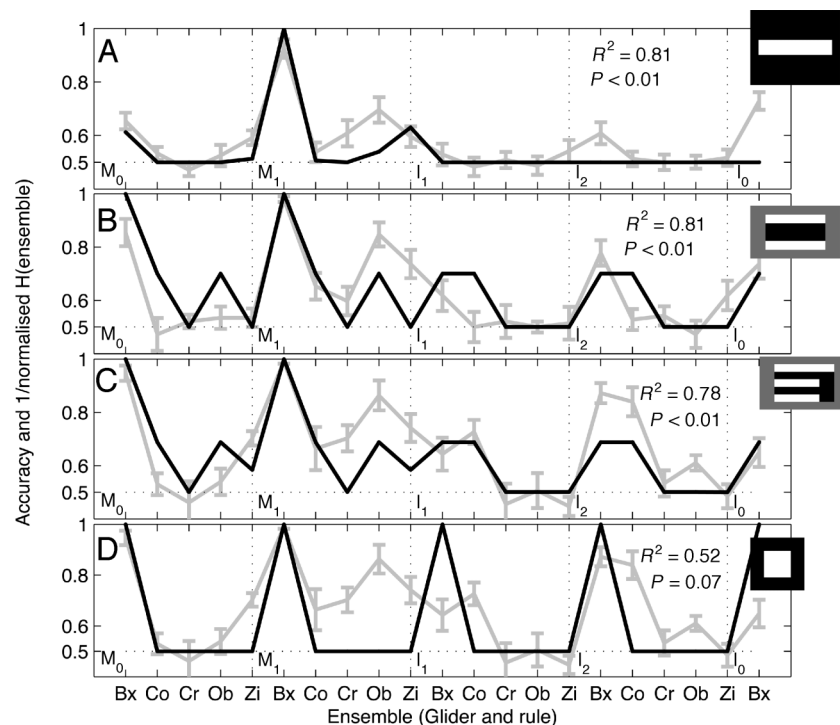


Figure 7. Best matches between 1/normalized based entropy measures (black) and PFs (gray) with corresponding R^2 and P values. The samplers used to generate each entropy function are shown at right of each plot. Preattentive data were best matched by samplers comprised of 1 or 2 strips for inexperienced (A) and experienced (B) data, respectively. For the attentive experienced data set, the best sampler comprised a three strip sampler (C). (D) A 2×2 pixel sampler distinguishes all Box textures from random, though completely misses structure in the other ensembles.

samplers comprised of three strips containing a total of 10 pixels (though almost the same correlation may be reached with 9) (Figure 7C, black). Using similarly shaped samplers containing more pixels (not shown) significantly decreased the fit to psychometric data. The high similarity between the attentive and preattentive PFs meant that the best matching samplers for each were also good models of the alternative data set. In general, for a sampler to generate a function that was reasonably correlated with a PF, it needed to contain a horizontal or vertical strip-like domain (Figure 8) comprised of more than 4 pixels. Changing the pixel distance between bars of input pixels within samplers decreased the similarity between measured information and the attentive PF (not shown). A critical reason strip-like samplers performed so well (Figures 7A, 7B, 7C, and 8) is that, as with human observers, they failed to differentiate a number of ensembles from random (most notably Figure 7A). A reason for employing almost the full range of ensembles was to examine not only the type of visual structure that is detectable, but also that which lies outside detection.

Figure 7D examines a 2×2 sampler based on an extensively studied type of higher-order structure considered in previous texture isotrison/isodipole studies (Joseph, Victor, & Optican, 1997; Julesz et al., 1978; Victor & Conte, 2004, 2005). This sampling domain captures structure in all the Box ensembles extremely

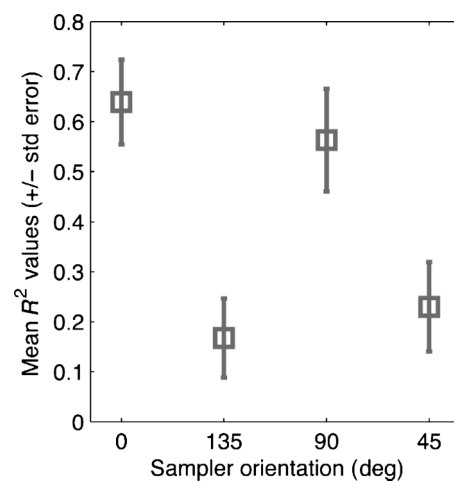


Figure 8. A bias for detecting vertical and horizontal higher-order structures. Averaged ($\pm$ SE) best matching R^2 values between psychometric functions and entropy functions derived from four sets of 15 samplers containing either horizontal (0°), vertical (90°), or oblique ($45^\circ/135^\circ$) input pixel arrangements. While each group differed in orientation, they were similar in terms of spatial arrangements and number of input pixels. For all samplers investigated, those containing horizontal or vertical pixel arrangements (0° , 90°) outperformed those containing oblique arrangements of input pixels.

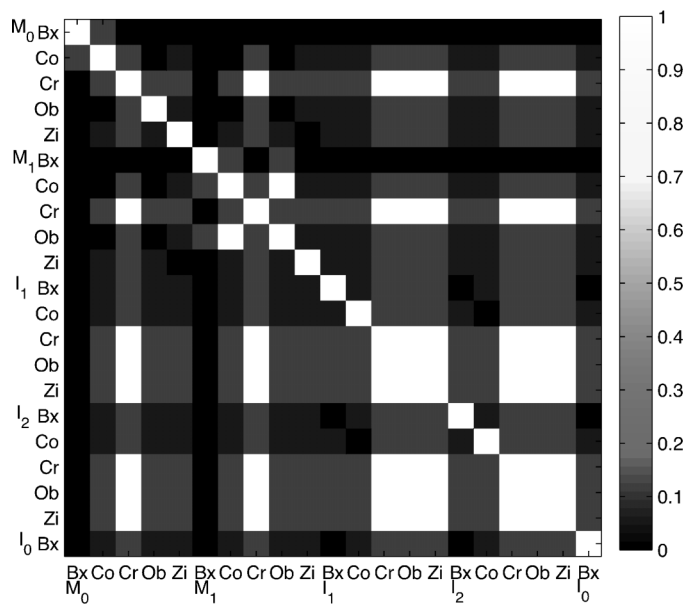


Figure 9. Measured similarity ($S(X, Y)$) between pairs of ensembles. Similarity is based on the number of words common to each pair of ensembles as found by the sampler that best matches the attentive psychometric data (see Figure 7C). Similarity values are between 0 (black = not at all similar) and 1 (white = identical). Highest entropy ensembles contain all possible words and are therefore predicted to appear similar (e.g., I_2 Cross, Oblong, and Zig Zag ensembles). In addition, some low entropy ensembles are also predicted to appear similar to one another. For example, the easily discriminated (highly structured) M_1 Box ensemble is predicted to look most like the M_1 Oblong and Corners ensembles. The less structured (less discriminable) M_1 Oblong and Corners ensembles are predicted to be highly similar, as are the M_1 Zig Zag, Cross, and M_0 Cross ensembles. These predictions may be examined by referring to Figures 1, 8, and 9.

well, though it does not distinguish between random patterns and any ensembles based on other gliders. The 2×2 sampler information function most closely resembles the attentive viewing PF, though the correlation is weak ($R^2 = 0.52$, $P = 0.07$).

By comparing how many words are shared between ensembles, it is possible to predict the degree to which they should appear similar to one another (provided the sampler entropy is a good match to visual processing). Described above, the similarity measure $S(X, Y)$ was employed to quantify such pairwise relationships. Figure 9 presents $S(X, Y)$ values (between 0 and 1) for all pairs of ensembles calculated using the sampler that best matched the experienced attentive data (see Figure 7C). Darker pixels reflect lower values and therefore lower degrees of measured similarity. For example, pixels along the diagonal in Figure 9 are white because they give the similarity between each ensemble and itself. Other white domains in Figure 9 mostly correspond to high entropy ensembles (e.g., I_0 and I_1 Cross, Oblong, and Zig Zag ensembles), that is, those containing all words that are

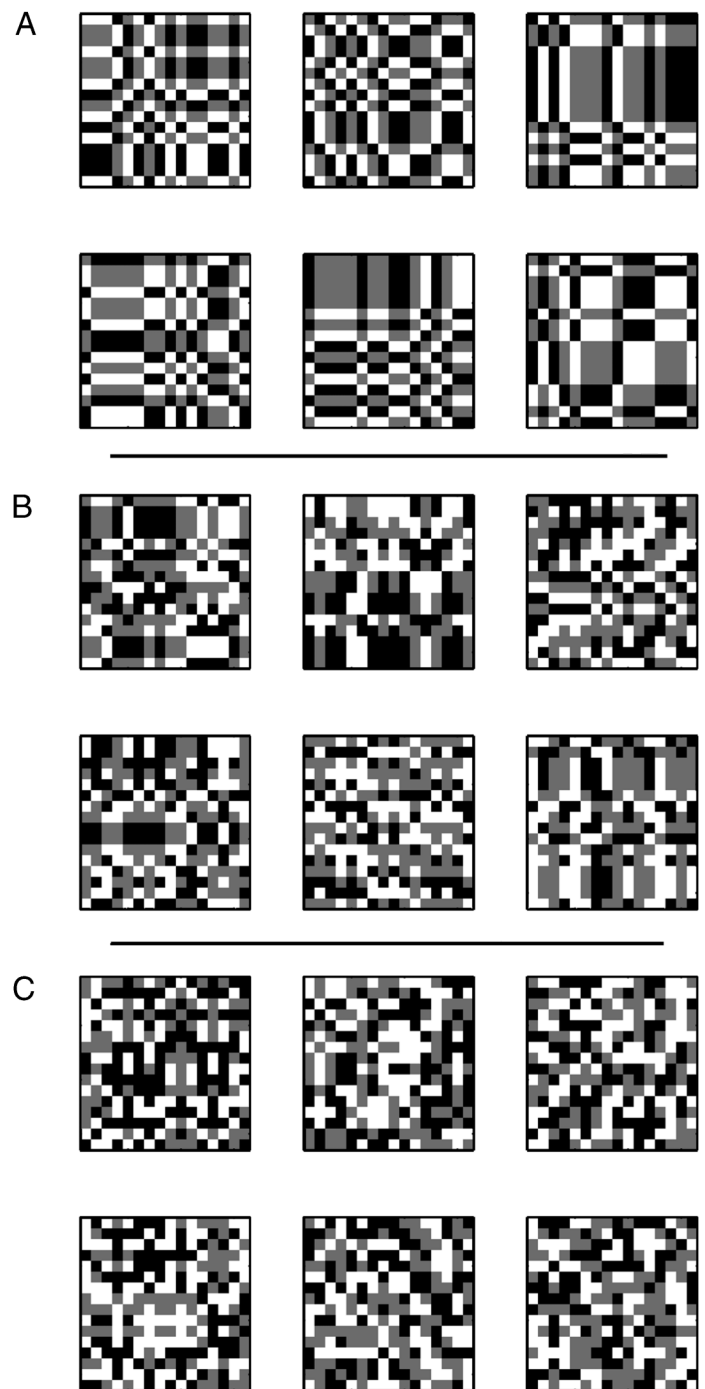


Figure 10. Examples of 3 lower entropy ensembles predicted to be similar based on their shared words (measured with the sampler that best matched the attentive data). (A) Six examples from the M_1 Box ensemble. (B) M_1 Oblong ensemble. (C) M_1 Corners ensemble. A and B appear somewhat similar, both having rectangular regions of a single contrast for example. B and C are predicted to be highly similar when attentively viewed. The visual similarity between these ensembles was confirmed psychophysically.

possible (in this case 59049) (see Figure 7C). Such textures appear unstructured (Figure 6) and therefore cannot be discriminated from one another, that is, they appear similar to each other and to noise patterns.

Interestingly, Figure 9 predicts that some more highly structured (low entropy) ensembles should also appear somewhat similar. For example, the easily discriminated M_1 Box ensemble is predicted to most closely resemble the M_1 Corners and Oblong ensembles (Figure 9). Six examples from each of these 3 ensembles are given in Figure 10. It can be seen that the I_1 Box (Figure 10A) and Oblong (10B) examples share features such as rectangular domains of a single contrast value. The M_1 Corners ensemble is predicted (Figure 9) to be highly similar to the M_1 Oblong ensemble. The relevant examples (B and C) in Figure 9 appear to share some features in common, and this similarity was studied psychophysically.

Eleven subjects from the above experiments completed 16 2AFC attentive (viewed for 30 seconds) trials in which

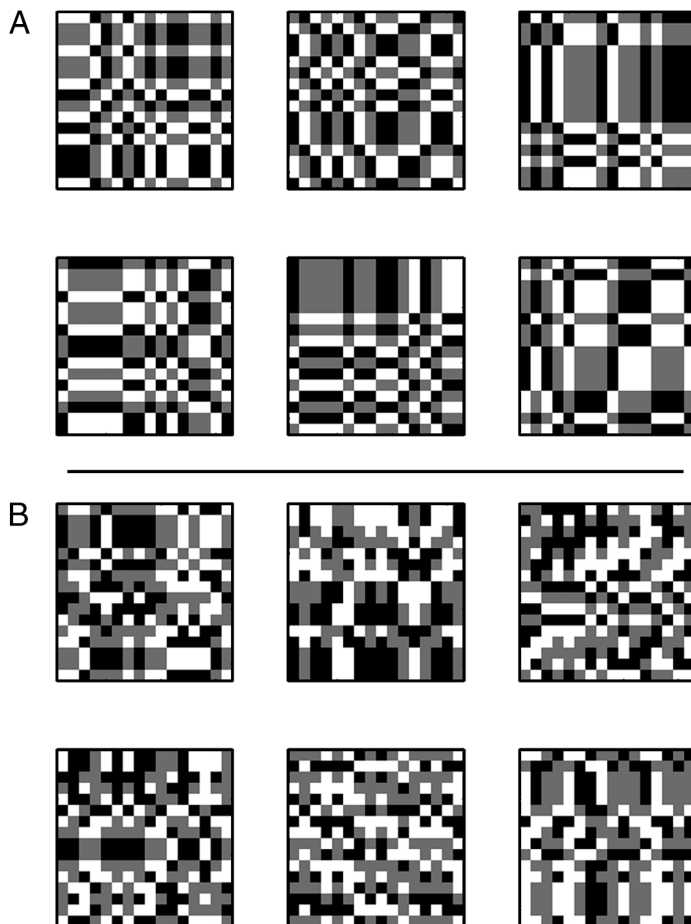


Figure 11. Examples of 2 lower entropy ensembles predicted to be similar based on their shared words (measured with the sampler that best matched the attentive data). (A) Six examples from the M_1 Zig Zag ensemble. (B) M_1 Cross ensemble. Perhaps the most noticeable similarity between A and B is that both contain obliquely orientated structures.

they had to discriminate between examples from M_1 Oblong and Corners ensembles. It was found that while the task was difficult, subjects can still discriminate between ensembles. Although the mean probability of correctly identifying an ensemble was low (mean probability = 0.6, standard deviation = 0.13), a *t*-test revealed that discrimination was significantly above chance level ($P = 0.03$). Nonetheless, the task is significantly harder than discriminating between M_1 Oblong examples and noise ($P < 0.001$). This suggests that some of the visual structures that differentiate the M_1 Oblong ensembles from noise cannot be used to differentiate between M_1 Oblong and Corners ensembles. It might therefore be that most of the perceived structure lies within horizontal parallel domains.

Figure 9 also predicts that the relatively structured (low entropy) M_1 Zig Zag and Cross ensembles should appear similar under attentive viewing conditions. Figure 11 gives six examples from these two ensembles. Perhaps the most noticeable similarity between the two ensembles is that both contain structures orientated at 45° . It is interesting that obliquely orientated structure may be captured by a sampler comprised of horizontal segments (Figure 4iii).

Discussion

The utility of statistically balanced texture ensembles has been apparent since Julesz and co-workers challenged essentially second-order (“filter-bank”) views of the striate cortex (V1) (Julesz, 1981; Julesz et al., 1978) (but see Zetzsche & Röhrbein, 2001). The battery of stimuli used in the present work has afforded a more controlled investigation of the interactions between image statistics and visual encoding than would have been possible with natural images or textures where accurate statistical categorization is difficult (e.g., Portilla & Simoncelli, 2000; Zhu, Wu, & Mumford, 1998). Using a large ternary battery of isotrigon ensembles that are isotropic below fourth order it has been possible to extensively examine a wide range of visually detectable and undetectable spatial structures. The abundance of these varied spatial elements places the present work, with its emphasis on learning characteristic structure, more in the domain of feature detection rather than texture processing. The above results may relate to early stages of this process, lying between representation in V1 and many of the highly nonlinear operations (feature extractions) that lead to feature specific representations such as face-selective cells for example (Bruce, Desimone, & Gross, 1981).

As previously demonstrated, the vast majority of complex spatial structure remains invisible to mechanisms that detect local form (Julesz, 1981; Julesz, Gilbert, Shep, & Frisch, 1973; Maddess & Nagai, 2001; Victor & Conte, 1991; Yellott, 1993). The above results imply that this insensitivity continues both with learning and attentive

processing. A comparison of best matching samplers suggests a reason that this may be the case. Namely, learning and attentive processing continue to modify a single strategy (the detection of relationships within single pixel wide straight lines) rather than search for alternative low dimensional spatial domains (e.g., Figure 7D) that might better recognize ensemble structure. Such persistence may point to physiological limits such as connectivity in the underlying detection mechanisms.

A simple way of understanding some of the constraints on the physiological substrates of local form detection is to consider a system that can uniquely distinguish all possible K th-order correlations within an N pixel image. To have a complete K th-order description, such a system would have to cast its input into a feature space. A feature space has $(N + K - 1)! / K!(N - 1)!$ dimensions described by a basis set comprised of K th-order products reflecting the order of the correlations being considered (Schölkopf, Smola, & Müller, 1998). For example, a second order description (e.g., a power spectrum) is based on all possible second-order products between each pixel in an image. For a two pixel image $\{a, b\}$, the space would be three dimensional with a basis $\{a^2, ab, b^2\}$. With increasing N or K , a “combinatorial explosion” quickly ensues, for example, a complete 4th order description of a four pixel image would require a 35 dimensional space. Although the complexity of computing such correlations may be reduced using kernel methods (Schölkopf et al., 1998), the task remains extremely complex.

For a mechanism encoding local structure over a small region, it may be better to restrict analysis to a subset of products (correlations) taken from a subset of spatial locations. The current approach is focused on the latter solution and gives no information about the types of correlations that may be involved. The results suggest that while the dimensionality may be stretched far enough to take into account at most 10 pixel interactions, these are confined to a narrow class of characteristic spatial domains. Modeling is currently being undertaken to determine how such local signals are combined (e.g., by simple summation or multiplication). It is not at all suggested the visual system compute anything like entropy, as such a system would be extremely complex (having to detect all ensemble words and their probabilities). Rather entropy (the log of the number of distinct words) has been used to point to regions over which information is extracted. Nonetheless, there could exist physiologically plausible mechanisms that sufficiently sub-sample the range of visual input so as to be driven by overall image complexity.

A potential limitation of the current work is that the data (PFs) may have been over-fitted by the large number of parameters employed. In one sense, the parameter space is as large as the number of possible words that may occur in a random pattern. There are several reasons to suppose the current findings are not the result of over-fitting. Firstly, the best performing samplers have predictive validity, in

that in addition to generating human-like entropy functions they also predict that some ensembles should appear similar to one another (Figure 9). For example, the visually structured M_1 Oblong and Corners examples are almost indistinguishable, as predicted by the number of words that are shared between them.

Importantly, only samplers containing strip-like horizontal or vertical elements produced closely matching ($R > 0.7$) functions (e.g., Figure 8). Such consistency would be highly unlikely if the match between entropy functions and data were simply the result of spurious correlations. Moreover, the present findings are supported by other empirical work and have some theoretical appeal (Victor & Conte, 1989, 1991, 1996). Successful models of V1 visual evoked potential (VEP) responses to interchanges between random and isodipole/isotrigon stimuli share critical features with the current findings (Victor & Conte, 1989, 1991). Such models are comprised of two nonlinear stages in which rectified responses of a number of linear high-pass filters (e.g., Gabor or edge) arrayed along a line are combined (accelerating non-linearity) to generate a local response to higher-order structure. Both the number of filters and the spatial extent over which their responses are combined are consistent with our information theoretic investigation.

The present results may help integrate such a model of spatial pooling with findings that demonstrate behavioral and physiological biases for detecting second-order stimuli (gratings) with horizontal or vertical ($0^\circ/90^\circ$) rather than oblique ($45^\circ/135^\circ$) orientations (Coppola, White, Fitzpatrick, & Purves, 1998; De Valois, Yund, & Hepler, 1982; Furmanski & Engel, 2000). If more second-order sensitive units (e.g., simple cells) are tuned to cardinal rather than oblique orientations (De Valois et al., 1982), then subsequent spatial pooling of these units may transfer the orientation bias for processing second-order structure to mechanisms sensitive to higher-order structure.

Currently we are investigating low parameter observational (principal component and independent component analysis) and two-stage nonlinear models of the texture discrimination task. In the case of observational models, it may be possible to uncover the number of mechanisms contributing to a subject's decision in the texture discrimination task. The extent to which each mechanism contributes to discrimination may then be computed, providing insight into to whether some ensembles are perceptually grouped (e.g., Figure 9). Both forms of modeling incorporate elements that may be directly related to the validity of the current results.

Conclusions

The perception of pattern or structure implies that the visual system detects ways in which the state of one region of an image may constrain states at other locations.

In this manner, the perception of structure corresponds to the detection of redundancy or lower information content. The current work suggests that an image is more likely to be perceived as structured or in the present context discriminable from noise, when redundancies (patterns) fall along narrow horizontal or vertical spatial domains. The scale-invariant spatial extent of such processing appears to take into account the values of at most 10 pixels, suggesting a computational limit to local form processing. By virtue of the shared statistical properties of texture ensembles, natural images (Chandler & Field, 2007) and image patches at the center of gaze (Krieger et al., 2000), such spatial biases and computational limits, are likely general features of both attentive and preattentive local form processing.

Acknowledgments

We acknowledge support from the ARC Centre of Excellence in Vision Science (CE0561903) and the ANU Centre for Visual Sciences.

Commercial relationships: none.

Corresponding author: Ryan R. L. Taylor.

Email: ryan.taylor@anu.edu.au.

Address: Centre for Visual Sciences, Research School of Biological Sciences, Australian National University, Canberra ACT 0200, Australia.

References

- Beason-Held, L. L., Purpura, K. P., Van Meter, J. W., Azari, N. P., Mangot, D. J., Optican, L. M., et al. (1998). PET reveals occipitotemporal pathway activation during elementary form perception in humans. *Visual Neuroscience*, 15, 503–510. [PubMed]
- Beck, J., Sutter, A., & Ivry, R. (1987). Spatial frequency channels and perceptual grouping in texture segregation. *Computer Vision Graphics Image Processing*, 37, 299–325.
- Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, 46, 369–384. [PubMed]
- Chandler, D. M., & Field, D. J. (2007). Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 24, 922–941. [PubMed]
- Coppola, D. M., White, L. E., Fitzpatrick, D., & Purves, D. (1998). Unequal representation of cardinal and oblique contours in ferret visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 2621–2623. [PubMed] [Article]
- Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley-Interscience.
- De Valois, R. L., Yund, E. W., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22, 531–544. [PubMed]
- Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local “association field.” *Vision Research*, 33, 173–193. [PubMed]
- Franz, M., & Schölkopf, B. (2006). *Implicit Volterra and Wiener series for higher-order image analysis*. Paper presented at the Advances in data analysis 30th Annual Conference.
- Furmanski, C. S., & Engel, S. A. (2000). An oblique effect in human primary visual cortex. *Nature Neuroscience*, 3, 535–536. [PubMed]
- Gilbert, E. N. (1980). Random colorings of a lattice on squares in the plane. *SIAM Journal on Algebraic and Discrete Methods*, 1, 152–159.
- Joseph, J. S., Victor, J. D., & Optican, L. M. (1997). Scaling effects in the perception of higher-order spatial correlations. *Vision Research*, 37, 3097–3107. [PubMed]
- Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, 290, 91–97. [PubMed]
- Julesz, B., Gilbert, E. N., Shep, L. A., & Frisch, H. L. (1973). Inability of humans to discriminate between visual textures that agree in second-order statistics—Revisited. *Perception*, 2, 391–405. [PubMed]
- Julesz, B., Gilbert, E. N., & Victor, J. D. (1978). Visual discrimination of textures with identical third-order statistics. *Biological Cybernetics*, 31, 137–140. [PubMed]
- Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, 13, 201–214. [PubMed]
- Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research*, 31, 679–691. [PubMed]
- Maddess, T., & Nagai, Y. (2001). Discriminating isotrigon textures [corrected]. *Vision Research*, 41, 3837–3860. [PubMed]
- Maddess, T., Nagai, Y., James, A. C., & Ankiewicz, A. (2004). Binary and ternary textures containing higher-order spatial correlations. *Vision Research*, 44, 1093–1113. [PubMed]

- Maddess, T., Nagai, Y., Victor, J. D., & Taylor, R. R. (2007). Multilevel isotricon textures. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 24, 278–293. [PubMed]
- Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40, 49–71.
- Purpura, K. P., Victor, J. D., & Katz, E. (1994). Striate cortex extracts higher-order spatial correlations from visual textures. *Proceedings of the National Academy of Sciences of the United States of America*, 91, 8482–8486. [PubMed] [Article]
- Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10, 1299–1319.
- Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24, 1193–1216. [PubMed]
- Victor, J. D. (1994). Images, statistics and textures: Implications of triple correlation uniqueness for texture statistics and the Julesz conjecture: Comment. *Journal of the Optical Society of America A*, 11, 1680–1684.
- Victor, J. D., Conte, M., Purpura, K., & Katz, E. (1995). Isodipole textures: A window on cortical mechanisms of form processing. In T. Papathomas, C. Chubb, A. Gorea, & E. Kowler (Eds.), *Early vision and beyond* (pp. 99–107). Cambridge: MIT Press.
- Victor, J. D., & Conte, M. M. (1989). Cortical interactions in texture processing: Scale and dynamics. *Visual Neuroscience*, 2, 297–313. [PubMed]
- Victor, J. D., & Conte, M. M. (1991). Spatial organization of nonlinear interactions in form perception. *Vision Research*, 31, 1457–1488. [PubMed]
- Victor, J. D., & Conte, M. M. (1996). The role of high-order phase correlations in texture processing. *Vision Research*, 36, 1615–1631. [PubMed]
- Victor, J. D., & Conte, M. M. (2004). Visual working memory for image statistics. *Vision Research*, 44, 541–556. [PubMed]
- Victor, J. D., & Conte, M. M. (2005). Local processes and spatial pooling in texture and symmetry detection. *Vision Research*, 45, 1063–1073. [PubMed]
- Victor, J. D., & Mast, J. (1991). A new statistic for steady-state evoked potentials. *Electroencephalography and Clinical Neurophysiology*, 78, 378–388. [PubMed]
- von der Heydt, R., Peterhans, E., & Dürsteler, M. R. (1992). Periodic-pattern-selective cells in monkey visual cortex. *Journal of Neuroscience*, 12, 1416–1434. [PubMed] [Article]
- Yellott, J. I. (1993). Implications of triple correlation uniqueness for texture statistics and the Julesz conjecture. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 10, 777–793.
- Zetzsche, C., & Röhrbein, F. (2001). Nonlinear and extra-classical receptive field properties and the statistics of natural scenes. *Network*, 12, 331–350. [PubMed]
- Zhu, S. C., Wu, Y. N., & Mumford, D. (1998). Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27, 107–126.