To appear in *Cold Spring Harbor Symposia, Volume 79 (2015): Cognition*

Representation of naturalistic image structure in the primate visual cortex

J. Anthony Movshon and Eero P. Simoncelli[1]

Center for Neural Science and [1]Howard Hughes Medical Institute,
New York University, New York, NY 10003, USA
Correspondence: movshon@nyu.edu

Abstract. The perception of complex visual patterns emerges from neuronal activity in a cascade of areas in the primate cerebral cortex. We have probed the early stages of this cascade with "naturalistic" texture stimuli designed to capture key statistical features of natural images. The responses of neurons in the primary visual cortex, V1, are relatively insensitive to the statistical information in these textures. However, in the area immediately downstream, V2, cells respond more vigorously to these stimuli than to matched control stimuli. Humans show BOLD fMRI responses in V1 and V2 that are consistent with the neuronal measurements in macaque. These fMRI measurements, as well as neurophysiological work by others, shows that true natural scenes become a more prominent driving feature of cortex downstream of V2. These results suggest a framework for thinking about how information about elementary visual features is transformed into the specific representations of scenes and objects found in areas higher in the visual pathway.

The framework: cascaded visual processing

In primates, the perception of complex visual patterns and objects emerges from neural activity as it is transformed through a cascade of areas in the cerebral cortex. Neurons in the primary visual cortex (V1) are selective for local orientation and spatial scale of visual input (Hubel and Wiesel, 1962, 1968; DeValois et al., 1982). Downstream areas contain neurons selective for more complex attributes, and this is presumably achieved by assembling particular combinations of their upstream afferents. But these attributes have proven difficult to discover with either perceptual or physiological measurements.

Given the ubiquity of orientation selectivity in primary visual cortex (Priebe and Ferster, 2012), it is intuitively appealing to assume that its computational purpose is to represent the the local orientation of edges. Over the past 50 years, the dominant view in both the computational and biological vision communities is that later stages of processing should somehow combine these local edge elements to construct corners, junctions, and more extensive contours, eventually leading to shapes, forms, and objects (Marr, 1982). Until recently, most computational research on object recognition was built around this paradigm, as well as much of the study of mid-level pattern perception, and physiological measurements in areas V2 and V4 of the ventral stream.

The intuitive appeal of the "edge paradigm" is partly due to its constructive nature. We imagine the visual system should analyze a visual scene much the way we would draw a picture of it. But this cartoon reasoning should be viewed with suspicion: the act of re-creating a scene with a pencil does not necessarily reveal the processes by which the scene was analyzed by the visual system. In fact, edges and contours make up a very small portion of most naturally-occurring visual scenes, and generating realistic drawings relies crucially on the introduction of additional elements such as shading and texture that are less easily described as the assembly of individual strokes. Moreover, 50 years of effort seem not to have brought us closer to an understanding of form vision.

An alternative (but not exclusive) minority view has coexisted with the edge-based view. In brief, the concept is that the visual system is more concerned with the representation of the "stuff" that lies between the edges, and less concerned with the edges themselves (Adelson and Bergen, 1991). In order to make this more concrete for the present discussion, let us focus on the specific case of visual texture.

Visual texture: models and human perception

"Visual texture" refers to portions of an image that are filled with repeated elements, often subject to some randomization in their location, size, color, orientation, etc; for example, an image of leaves, or pebbles, or tree bark (Fig. 1a). Lettvin (1976), offered this insight: "Let us say that to the extent that visible objects are different and far apart, they are *forms*. To the extent that they are similar and congregated they are a *texture*. A man has form; a crowd has man-texture. A leaf has form; an arbor has leaf-texture, and so on." Bela Julesz pioneered the statistical characterization of visual texture by proposing that the statistics of image pixels, measured up to some order, should suffice to partition textures into classes that are indistinguishable to a human observer (Julesz, 1962). Implicitly, Julesz was asserting that the human visual system represented visual texture by measuring these statistics, and only these statistics. In this case, the theory predicts that any two images that are identical in this statistical sense must appear identical to a human.

Critical to Julesz' framing of the problem was the desire for a *minimal* set of statistics, as well as the plan to experimentally validate the model by seeking perceptual counter-examples. Julesz and colleagues initially thought that pairwise statistics were sufficient (Julesz et al., 1973), but then disproved this by producing hand-constructed example pairs of textures with identical statistics through second (and even third) order that were easily distinguished by human observers (Caelli and Julesz, 1978; Julesz et al., 1978). Given the falsification of this particular instantiation of his statistical theory, Julesz abandoned the approach altogether, and began to develop a constructive theory of texture based on randomly placed "texton" features. Like the "edge paradigm", this was appealing for its constructive nature, and easily led to stimuli and experiments, but proved far more difficult to interpret in terms of perceptual or physiological representation.

Julesz' statistical conceptualization was sound, but his definition of the statistical model in terms of the order of pixel statistics was problematic. Increasing statistical order leads to ever more nonlinear interactions of multiple pixels, with no restriction on the spatial extent of those interactions. Responses of early visual neurons, on the other hand, are generally described in terms of simple combinations of *local* image intensity. For example, retinal ganglion responses are commonly described as the rectified response of a center-surround linear filter. V1 simple cell responses are commonly described as rectified (or rectified and squared) responses of oriented linear filters (Heeger, 1992). And V1 complex cell responses are described as an average over these simple cell responses, all having the same orientation preference, but differing in the spatial location of their filters (Hubel & Wiesel, 1962; Movshon et al., 1978; Adelson and Bergen, 1985).

Measuring statistics in a physiologically consistent format can lead to much more powerful model of texture. Specifically, a model for texture based on the pairwise correlations between model simple- and complex-cell responses, at nearby positions, orientations, and scales (Fig 1d) can capture many of the salient features of natural texture images (Portilla and Simoncelli, 2000). The correlations are computed by averaging the product of pairs of responses over the spatial extent of the image. To demonstrate that this model captures the visually relevant attributes of a given texture, the authors followed Julesz' paradigm, synthesizing new images with the same model responses, and then asking whether the resulting images were similar in appearance to the original. In brief: one initializes an image with white noise, and then adjusts the pixels according to the gradient of the squared error of the model responses, relative to the desired model responses. This procedure converges after a relatively small number of iterations, and produces images that are similar in appearance to the original (Portilla and Simoncelli, 2000; Balas, 2011) (Fig. 1e). It seems remarkable that responses of these simplistic models (cascades of filtering, rectification, pooling), which provide only a crude statistical summary of the image content, are sufficient to capture such sophisticated features. But it is worth noting that a reduced model which uses only the averages of V1 responses, computed for cells at each orientation and size (Fig. 1b) and therefore measuring only spectral information, is insufficient to capture the features of most naturally-occurring textures (Fig. 1c)

Given that synthesized textures capture much of the visual appearance of the textures from which their statistics are derived, it is natural to ask whether (and where) these statistics might be represented in the brain. While the texture model described in the previous paragraph was not originally intended as a model for post-V1 physiology, two modifications allow it to be interpreted as such, at the abstract level of population representation. First, the statistics can be gathered *locally*, over regions corresponding to receptive fields in a visual area downstream of V1. A natural candidate is V2, which receives strong direct input from V1, and whose neurons have receptive fields that are roughly twice the size of those in V1 at any given eccentricity (Gattass et al., 1981). As in most visual areas, receptive field sizes in V2 increase roughly in proportion to eccentricity. Second, the products of V1-like afferents that are averaged when computing model correlations can be replaced with the squares of summed afferents (Adelson and Bergen, 1985). The product is implicitly represented (it is the "cross term"), and two textures with the same correlations will also have the same values in this modified model. More importantly, this modified model assumes the same form as models of V1 complex cells (a sum of squared linear filter responses), and thus instantiates a cascade model of cortical computation in which the same elementary operations are performed in each cortical area, differing only in the inputs they receive.

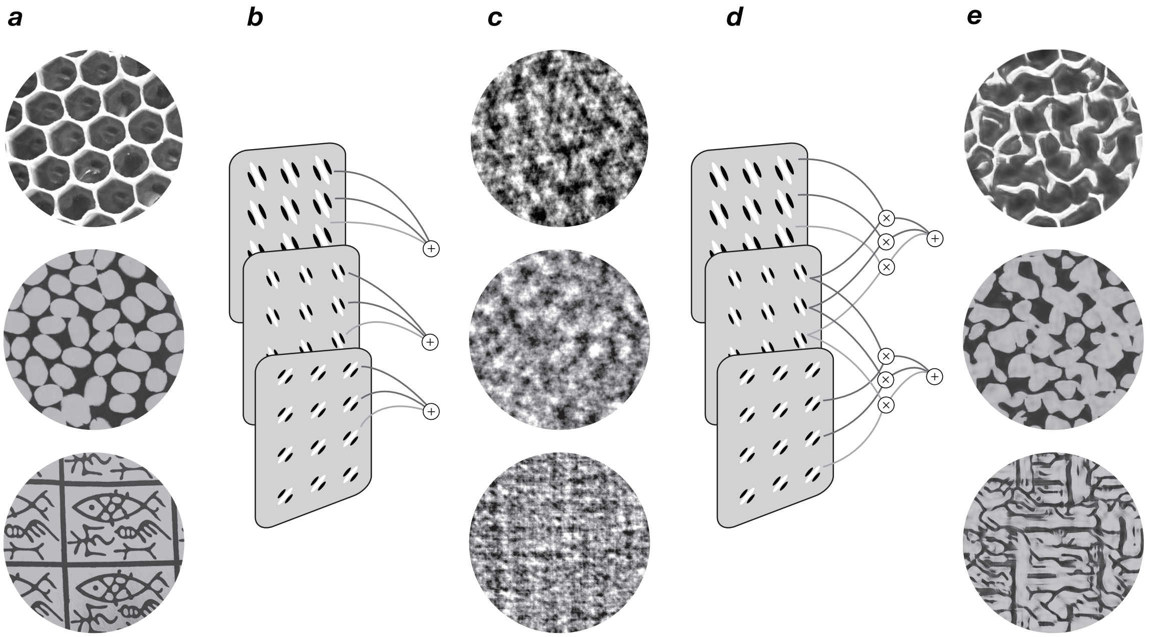

Figure 1. Naturalistic textures, and synthetic images with matching statistics. a. Photographs of three different textures. **b.** Spectral statistics, as captured with a V1-like model. Statistics are computed as spatial averages over responses of arrays of model neurons, whose receptive fields have a particular preferred orientation and scale. **c.** Images synthesized so as to have identical spectral statistics to the corresponding original images (column a). **d.** Joint (correlation) statistics of a V1-like model. Statistics are computed as correlations (average of pairwise products) both within and across arrays of model V1 neurons (Portilla and Simoncelli, 2000). **e.** Images synthesized to have identical joint statistics.

This physiological version of the texture model, in which statistics are computed over regions roughly the size of V2 receptive fields, can be used to synthesize images whose local statistics over each region are matched to those of an existing photograph. For a human observer whose gaze is directed to the correct location in the image (the sizes of model statistical regions are chosen to match those of V2 receptive fields only for a particular center of gaze), these synthetic images are indistinguishable from the original photograph, despite distortions whose severity increases with eccentricity (Fig. 2) (Freeman and Simoncelli, 2011). This provides a direct demonstration that the information discarded by the model, which retains only a crude statistical summary of the content within each local spatial region, is also discarded by the human visual system. This also provides an explanation for the visual phenomenon of “crowding” (Bouma, 1970), in which humans fail to recognize peripherally presented objects that are surrounded by background clutter (Balas, et al., 2009; Freeman and Simoncelli, 2011).

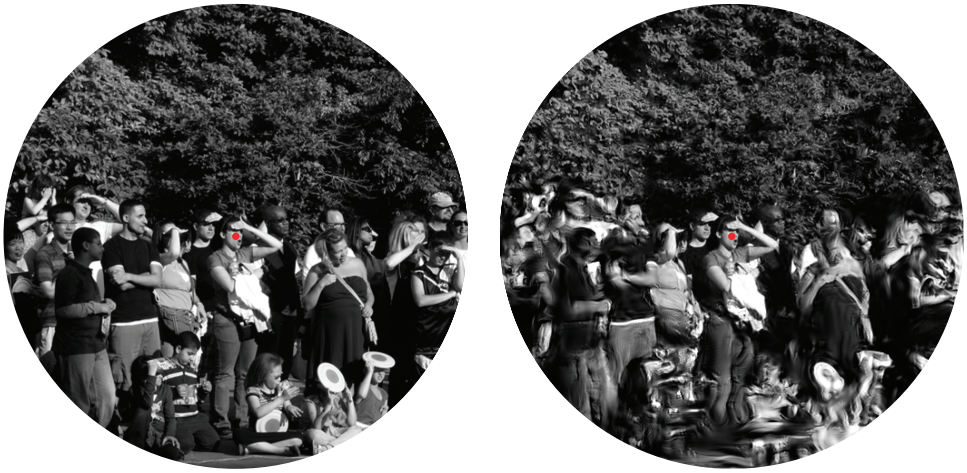

Figure 2. Example photograph and a random image synthesized to have matching texture statistics over spatial regions the size of V2 receptive fields. V2 receptive fields are roughly twice the diameter of V1 receptive fields, and grow approximately linearly with distance from the fovea. When viewed in alternation by human subjects with their gaze fixated on the image center (red dot), the two images are virtually indistinguishable, despite dramatic distortions in the periphery. Thus the statistics of the model, although they provide only a partial summary of the original scene, are sufficient to capture its visual appearance.

Armed with the hypothesis that V2 represents local statistics of the sort used in this statistical texture model, a natural next step is to examine the physiological responses of V2 neurons to these stimuli.

Macaque cortical unit responses to naturalistic texture

It has proved difficult to assign specific visual functions to V2, or at least to deduce them from analysis of V2 neuronal responses to simple parametric stimuli, because these are qualitatively quite similar to responses measured in V1. V2 cells are commonly selective for orientation, direction, spatial frequency, drift rate or speed, color, binocular disparity – all the dimensions for which V1 neurons are selective (Hubel and Wiesel, 1965; Hubel and Livingstone, 1987; Levitt et al., 1994; Hegde and Van Essen, 2007). There are quantitative differences. An idea that has its roots in early work of Hubel and Wiesel (1965) is that cells in V2 are selective for simple extensions of the basic contour responses measured in V1 – selectivity has been reported for such elaborated features as curvature, angle, convexity, and illusory borders (von der Heydt and Peterhans, 1989; Zhou et al., 2000; Hegde and Van Essen, 2000; Ito and Komatsu, 2004; Anzai et al., 2007; Willmore et al., 2010). However, while some cells show selectivity for these various attributes, many do not, and when closely examined population selectivity may not differ substantially from that seen when comparable measurements are made in V1 (Hegde and Van Essen, 2007).

Based on the findings described in the previous section, we hypothesized that cells in V2, but not those in V1, might represent the spatial correlations that distinguish naturalistic texture from spectrally-matched control stimuli. V2 receives a strong direct input from V1, and the size of its receptive fields match human thresholds for discriminating scrambled textures (Freeman and Simoncelli, 2011; Fig. 2). We therefore decided to begin with a comparison of response properties of cells in V1 and V2, using a set of textures chosen to span a reasonable range of naturally occurring variation in simple image statistics.

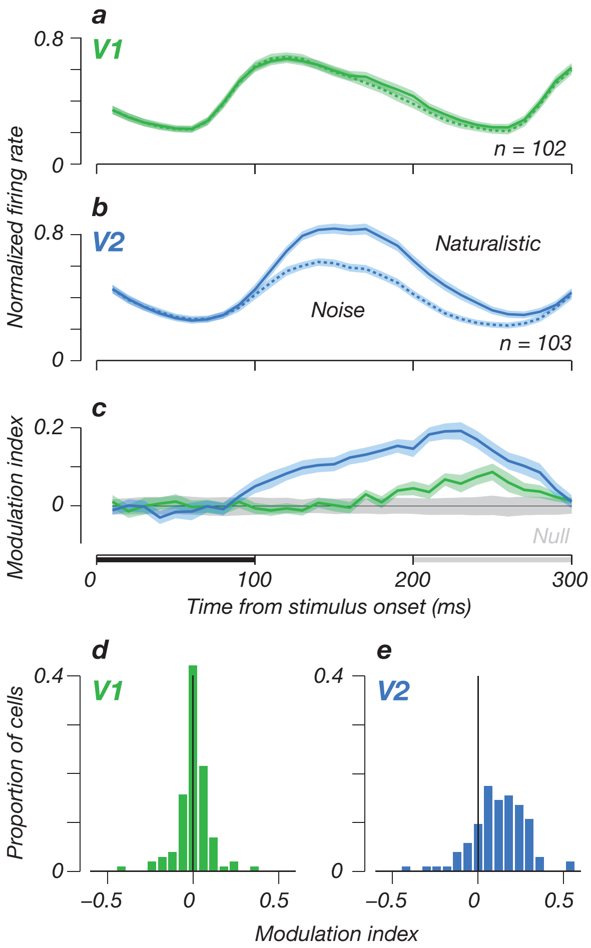

Figure 3. Responses of neurons in macaque V1 ad V2 to naturalistic textures and spectrally-matched noise controls. We measured the responses of isolated cells in V1 and V2 cells to a randomly-interleaved sequence of naturalistic textures and spectrally-matched noise. We presented the textures for 100 ms in a 4 deg window centered on the receptive field, with a 100 ms blank interval between stimuli. **a.** Normalized mean responses of 102 cells from V1. Solid line – naturalistic stimuli, dashed line – noise controls. Stimulus time course indicated at bottom. **b.** Normalized mean responses of 103 cells from V2; conventions as in a. **c.** A modulation index, computed as the difference between naturalistic and noise responses divided by their sum, plotted for V1 and V2 populations. Shaded bands in a, b, and c indicate 95% confidence intervals. **d.** Distribution of the modulation index for 102 cells in V1. **e.** Distribution of the modulation index for 103 cells in V2. After Freeman et al. (2013a).

The results of our neurophysiological texture experiments are summarized (Freeman et al., 2013a)in Fig. 3. In Fig. 3a and 3b, we show the average response time courses for populations of about 100 units recorded from V1 (Fig. 3a) and V2 (Fig. 3b). In V1, responses to naturalistic textures and spectrally matched noise controls were almost indistinguishable, while in V2 responses to naturalistic textures were robustly larger than responses to noise. The difference is captured in Fig 3c with a "modulation index", which shows strong modulation for the V2 population over the entire time course of response, and only very weak modulation for the V1 population, emerging very late in the response. The average traces do not reveal how consistent the responses were for different neurons, so Fig. 3d and 3e show the distribution of the modulation index for neuronal populations in V1 and V2. The V1 data show little dispersion about a mean near 0, while the mean for V2 was substantially and significantly shifted to positive values.

Human cortical responses to naturalistic texture and natural images

Given the robustly enhanced neuronal responses to naturalistic textures that we observed in unit responses recorded from V2, we decided to look for evidence of similar responses in human visual cortex, using the technique of BOLD fMRI (Freeman et al., 2013b). We measured responses in three subjects to an alternating sequence of spectrally-matched control stimuli and naturalistic textures (Fig 4a), presented in an annular region surrounding the fixation target. We render the responses on a flattened map of the posterior pole of the cortex in a representative subject; the marked borders between visual areas were established in earlier topographic mapping studies using standard techniques (Wandell et al 2007; Gardner et al., 2008). The alternation of naturalistic and noise stimuli notably failed to modulate the BOLD signal in V1, but in V2 and adjacent areas, the response was robust and reliable. This is entirely consistent with our measurements of neuronal responses in macaque V1 and V2 (Freeman et al., 2013a).

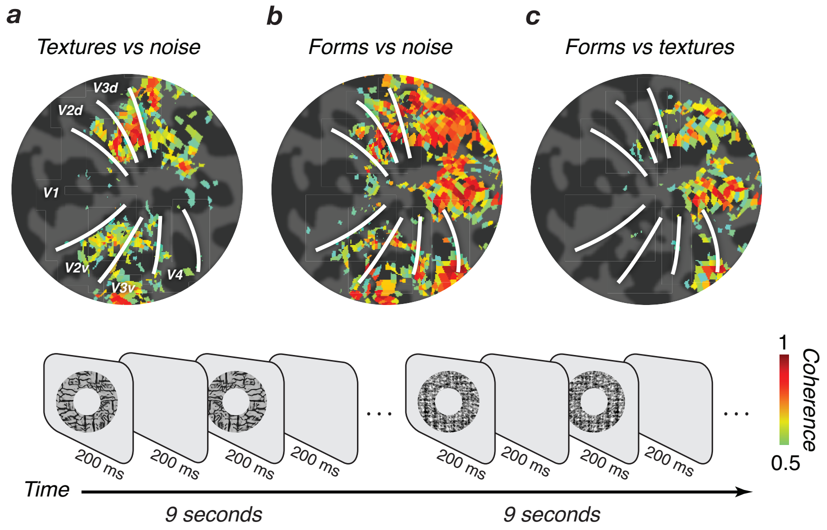

Figure 4. Responses of human visual cortex to natural images, naturalistic textures, and spectrally-matched noise controls. Flattened maps of the posterior pole of the right hemisphere of a representative observer showing modulation of BOLD fMRI responses to alternating sequences of stimuli in a block design whose time course is schematically diagrammed at the bottom: briefly presented repeated stimuli (200 ms on, 200 ms off) were presented in 9 s blocks, alternating between the conditions being compared. Area boundaries are derived from a separate topographic mapping experiment. In each panel, the pseudocolor scale represents the coherence of the BOLD response component synchronized with the time course of the stimulus exchange. **a.** Alternation between naturalistic textures and spectrally matched noise controls. **b.** Alternation between natural images and spectrally matched noise controls. **c.** Alternation between natural images and naturalistic textures. From Freeman et al. (2013b).

The broad view of the brain given by functional imaging allowed us also to explore the way in which image representation begins to move beyond the representation of the kind of naturalistic image statistics that are effective in driving V2; in particular, we have begun to ask whether further elaborations of the cascade of visual processing leads downstream areas to respond differentially not only to images with naturalistic statistics, but also to natural images themselves. Physiological measurements of unit responses in macaque extrastriate areas V4 and IT by Rust and DiCarlo (2010) suggest this possibility, showing that units there respond better to photographs of natural scenes than to scrambled photographs made to match the statistics of the texture model. Using the same block design, we compared fMRI activation in human visual cortex to natural images with those to spectrally matched noise (Fig 4b), and also compared activation by natural images and naturalistic textures derived from those images (Fig. 4c)(Freeman et al., 2013b). While response modulation in V2 and V3 was similar for the exchange of natural images with noise, response modulation in areas anterior to V3 was stronger than when naturalistic textures were exchanged with noise (Fig. 4b). As one might then expect, the exchange of natural images for naturalistic textures evoked little response in early areas (V1-V3), but strongly activated anterior areas (V4 and other downstream areas). In aggregate, these preliminary measurements suggest that areas downstream of V2 and V3 continue the process of elaboration begun in V2, and effect at least part of the job of transforming representations of the statistical features of images into representations of the particular features of objects and scenes that are found in the visual areas of the inferior temporal lobe (DiCarlo et al., 2012)

Conclusion and outlook

We have established that V2 reformats the relatively simple visual representation provided by V1 so as to make *explicit* relationships of activity that are *implicit* in the input population. This is akin to the computation of orientation selectivity from non-oriented inputs that is executed by circuits in V1 – the orientational structure of the visual image is implicit in the responses of LGN neurons that provide input to V1, but is made explicit in V1 (Hubel and Wiesel, 1962; Priebe and Ferster, 2012). Another example of such cascaded computation is the selectivity for pattern motion seen in directionally selective neurons in area MT – these neurons respond to complex object motion in a way that is implied, but not directly represented in the inputs from V1 (Movshon et al., 1985; Simoncelli and Heeger, 1998; Rust et al., 2006). This recurring motif of cascaded computation is likely the basis of other forms of selectivity in the visual pathway, such as the selective responses to forms and objects observed in V4 and IT (Brincat and Connor, 2004; Serre et al., 2007; Rust and DiCarlo, 2010). Such elaboration is also likely the basis of our finding that V4 and other areas downstream of V2 seem to respond preferentially to natural images of objects and scenes, compared to statistically matched textures derived from those images (Freeman et al., 2013b; Fig. 4c).

What can we say about the responses of individual V2 neurons on the basis of our analysis of group data? The population of neurons in V1 can represent orientation for purposes, say, of discrimination (Graf et al., 2011). But this is not the same as being able to describe the responses of each V1 cell, for which a particular set of properties and circuits are in play (Priebe and Ferster, 2012). The same is true for V2 – knowing that the population of V2 neurons can represent statistical structure in images does not uniquely specify the response properties of individual V2 neurons, and a major challenge for ongoing research will be to create such models. Related work in V4 has explored computations that produce some forms of texture specificity (Okazawa et al., 2015), and we have begun to explore models to account for the texture responses of V2 neurons. The structure of the statistical model from which the texture stimuli are generated (Portilla and Simoncelli, 2000) provides a starting point. This model combines particular V1-like responses by correlation – an average of pairwise products – a calculation for which a neural implementation has not been demonstrated in mammals (though see Jones and Gabbiani, 2012 for an invertebrate example). As shown by Adelson and Bergen (1985), an equivalent model can be constructed by squaring the sums of these pairs, creating an alternative neural circuit implementation that may be more plausible for mammalian cortex. Such models make testable predictions (see Emerson et al., 1992), and fitting and testing V2 models of this form against data from individual neurons is an important next step.

In addition, neural network models that are optimized for performance of visual recognition may provide some constraints (Serre et al. 2007; Kaligh-Razavi and Kriegeskorte, 2014; Yamins et al, 2014). In the past few years, hierarchical neural network models, optimized for recognition performance on large image databases, have outstripped previous models working within the classical edge-based paradigm. The success of these networks is still mysterious, and the details of their learned internal representations have not been carefully studied (Kaligh-Razavi and Kriegeskorte, 2014; but see Zeiler and Fergus 2014). The design of these networks is often inspired by physiology (Jarrett et al., 2009), and given the similarity to the construction of the texture model described here, are likely to capture similar texture-like attributes in intermediate stages. Images synthesized from these representations may therefore prove useful as stimuli in perceptual or physiological investigations, and the precise form of internal representation in these networks may provide inspiration for the design of physiological response models.

Our exploration of texture representations in cortex was driven by perceptual phenomena and mechanisms. Now we can begin to ask how to relate cortical neural circuits to those perceptual phenomena. We have found that the average modulation index of V2 neurons for particular texture images is correlated with the ability of human observers to detect the presence of the model statistics in synthetic versions of those images (Freeman et al., 2013a). In the context of a more natural perceptual task, we have obtained preliminary evidence that V2 neurons also provide a better substrate for classifying textures into categories than V1 neurons – as a population, V2 neurons are selective for the properties of particular textures, while being tolerant of image variations that preserve texture identity, such as those that arise during the synthesis procedure (Ziemba et al., 2012). In this regard, V2 responses to texture are analogous to IT responses to objects and forms – selective for particular forms and at the same time robust to transformations that preserve the identity of those forms (Rust and DiCarlo, 2010).

A central question is whether the form and object selectivity of these later stages is constructed from texture-selective neurons in area V2, or whether it is instead arises through some other form-specific pathway. Lettvin (1976) considers the relationship between form and texture: “One can imagine shapes spatially interfering with each other to comprise texture; or else suppose that texture is primitive and that textures combine to produce forms – just as letters combine to make words.”. Lettvin himself comes down on the second side, arguing that texture is primitive and form more highly evolved. This conjecture is compatible with our observations that mechanisms in V2, relatively early in the visual pathway, respond invariantly to texture. Only later, in areas like the inferotemporal cortex, do responses become selective and invariant for objects and scenes.

Acknowledgements

The work reviewed in this paper was done in collaboration with Javier Portilla, Jeremy Freeman, Corey Ziemba, Michael Landy, and David Heeger, whose many contributions we gratefully acknowledge. This work was supported by a grant from the NIH (R01 EY022428), and by the Howard Hughes Medical Institute.

References

Adelson EH, Bergen JR. 1985. Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A* **2:** 284-99.

Adelson EH, Bergen JR. 1991. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pp. 3-20, M S Landy and J A Movshon (eds). Cambridge, MA: MIT Press.

Anzai A, Peng X, Van Essen DC. 2007. Neurons in monkey visual area V2 encode combinations of orientations. *Nat Neurosci* **10:** 1313-21.

Balas B, Nakano L, Rosenholtz R. 2009. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis* **9:** 13-18.

Bouma H. 1970. Interaction effects in parafoveal letter recognition. *Nature* **226:** 177-178.

Brincat SL, Connor CE. 2004. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat Neurosci* **7:** 880-886.

Caelli T, Julesz B. 1978. On perceptual analyzers underlying visual texture discrimination: part I. *Biol Cybern* **28:** 167-175.

De Valois RL, Albrecht DG, Thorell LG. 1982. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Res* **22:** 545-559.

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? *Neuron* 7**3:** 415-434.

Emerson RC, Bergen JR, Adelson EH. 1992. Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Res* **32:** 203-218.

Freeman J, Simoncelli EP. 2011. Metamers of the ventral stream. *Nat Neurosci* **14:** 1195-1201.

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA. 2013a. A functional and perceptual signature of the second visual area in primates. *Nat Neurosci* **16:** 974-981.

Freeman J, Ziemba CM, Simoncelli EP, Movshon JA. 2013b. Functionally partitioning the ventral stream with controlled natural stimuli. Society for Neuroscience Abstracts 406.01, Online.

Gardner JL, Merriam EP, Movshon JA, Heeger DJ. 2008. Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. *The Journal of neuroscience* **28:** 3988-3999.

Gattass R, Gross CG, Sandell JH. 1981. Visual topography of V2 in the macaque. J *Comp Neurol* **201:** 519-539.

Graf AB, Kohn A, Jazayeri M, Movshon JA. 2011. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nat Neurosci* **14:** 239-245.

Heeger DJ. 1992. Half-squaring in responses of cat striate cells. *Vis Neurosci* **9:** 427-443.

Hegdé J, Van Essen DC. 2007. A comparative study of shape representation in macaque visual areas v2 and v4. *Cereb Cortex* **17:** 1100-1116.

Hegdé J, Van Essen DC. 2000. Selectivity for complex shapes in primate visual area V2. *J Neurosci* **20:** RC61.

Hubel DH, Livingstone MS. 1987. Segregation of form, color, and stereopsis in primate area 18. *J Neurosci* **7:** 3378-3415.

Hubel DH, Wiesel TN. 1968. Receptive fields and functional architecture of monkey striate cortex. *J Physiol* **195:** 215-43.

Hubel DH, Wiesel TN. 1965. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Neurophysiol* **28:** 229-289.

Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* **160:** 106-154.

Ito M, Komatsu H. 2004. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J Neurosci* **24:** 3313-3324.

Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y. 2009. What is the best multi-stage architecture for object recognition?. In *Computer Vision, 2009 IEEE 12th International Conference,* pp. 2146-2153. IEEE.

Jones PW, Gabbiani F. 2012. Impact of neural noise on a sensory-motor pathway signaling impending collision. *Journal of Neurophysiology* **107:** 1067-1079.

Julesz B. 1962. Visual pattern discrimination. *I.R.E. Trans. Information Theory* **IT-8**: 84-92.

Julesz B, Gilbert EN, Victor JD. 1978. Visual discrimination of textures with identical third-order statistics. *Biol Cybern* **31:** 137-140.

Julesz B, Gilbert EN, Shepp LA, Frisch HL. 1973. Inability of humans to discriminate between visual textures that agree in second-order statistics-revisited. *Perception* **2:** 391-405.

Khaligh-Razavi SM, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol* **10:** e1003915.

Lettvin JY. 1976. On seeing sidelong. *The Sciences* **16:** 10-20.

Levitt JB, Kiper DC, Movshon JA. 1994. Receptive fields and functional architecture of macaque V2. *J Neurophysiol* **71:** 2517-2542.

Marr D. 1982. Vision. San Francisco: W H Freeman.

Movshon JA, Thompson ID, Tolhurst DJ. 1978. Receptive field organization of complex cells in the cat's striate cortex. *J Physiol* **283**: 79–99

Movshon JA, Adelson EH, Gizzi MS, Newsome WT. 1985. The analysis of moving visual patterns. *Pattern recognition mechanisms* **54:** 117-151.

Okazawa G, Tajima S, Komatsu H. 2015. Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc Natl Acad Sci U S A* **112:** E351-E360.

Portilla J, Simoncelli EP. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision* **40:** 49-70.

Priebe NJ, Ferster D. 2012. Mechanisms of neuronal computation in mammalian visual cortex. *Neuron* **75:** 194-208.

Rust NC, Dicarlo JJ. 2010. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J Neurosci* **30:** 12978-12995.

Rust NC, Mante V, Simoncelli EP, Movshon JA. 2006. How MT cells analyze the motion of visual patterns. *Nat Neurosci* **9:** 1421-1431.

Serre T, Oliva A, Poggio T. 2007. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences* **104:** 6424-6429.

Simoncelli EP, Heeger DJ. 1998. A model of neuronal responses in visual area MT. *Vision Res* **38:** 743-761.

von der Heydt R, Peterhans E. 1989. Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *J Neurosci* **9:** 1731-1748.

Wandell BA, Dumoulin SO, Brewer AA. 2007. Visual field maps in human cortex. *Neuron* **56:** 366-383.

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci USA* **111:** 8619-8624.

Zeiler ND, Fergus R. 2014. Visualizing and understanding convolutional networks. *Computer Vision – ECCV* 2014, 816-833.

Zhou H, Friedman HS, Von Der Heydt R. 2000. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience* **20:** 6594-6611.

Ziemba CM, Freeman J, Movshon JA, Simoncelli EP. 2012. Selectivity and invariance are greater in macaque V2 than V1. CoSyNe Abstracts II-91. Online.