nature neuroscience

Metamers of the ventral stream

Jeremy Freeman[1] & Eero P Simoncelli[1–3]

The human capacity to recognize complex visual patterns emerges in a sequence of brain areas known as the ventral stream, beginning with primary visual cortex (V1). We developed a population model for mid-ventral processing, in which nonlinear combinations of V1 responses are averaged in receptive fields that grow with eccentricity. To test the model, we generated novel forms of visual metamers, stimuli that differ physically but look the same. We developed a behavioral protocol that uses metameric stimuli to estimate the receptive field sizes in which the model features are represented. Because receptive field sizes change along the ventral stream, our behavioral results can identify the visual area corresponding to the representation. Measurements in human observers implicate visual area V2, providing a new functional account of neurons in this area. The model also explains deficits of peripheral vision known as crowding, and provides a quantitative framework for assessing the capabilities and limitations of everyday vision.

© 2011 Nature America, Inc. All rights reserved.

The ventral visual stream is a series of cortical areas that represent spatial patterns, scenes and objects[1]. V1 is the earliest and most thoroughly characterized area. Individual V1 cells encode information about local orientation and spatial frequency[2], and simple computational models can describe neural responses as a function of visual input[3]. Substantial progress has also been made in understanding later stages, such as inferotemporal cortex, where neurons exhibit complex object-selective responses[4]. However, the transformations between V1 and inferotemporal cortex remain a mystery.

Several observations from physiology and theory can help to constrain the study of this problem. It has been shown that receptive field sizes increase along the ventral stream. Many models of visual pattern recognition[5–10] have proposed that increases in spatial pooling provide invariance to geometric transformations (for example, changes in position or size). In addition, it is well established that receptive field sizes scale linearly with eccentricity in individual areas and that this rate of scaling is larger in each successive area along the ventral stream, providing a signature that distinguishes different areas[11–13].

We hypothesize that the increase in spatial pooling, both in successive ventral stream areas and with eccentricity, induces an irretrievable loss of information. Stimuli that differ only in terms of this lost information will yield identical population-level responses. If the human observer is unable to access the discarded information, such stimuli will be perceptually indistinguishable; thus, we refer to them as metamers. Visual metamers were crucial to one of the earliest and most successful endeavors in vision science: the elucidation of human trichromacy. Behavioral experiments predicted the loss of spectral information in cone photoreceptors 100 years before the physiological mechanisms were confirmed[14]. The concept of metamerism is not limited to trichromacy, however, and a number of studies have used it to understand aspects of pattern or texture vision[15–17].

We developed a population-level functional model for ventral stream computation beyond V1 that allowed us to synthesize and examine the perception of a new type of visual metamer. The first stage of the model decomposes an image with a population of oriented V1-like receptive fields. The second stage computes local averages of nonlinear combinations of these responses over regions that scale in size linearly with eccentricity, according to a scaling constant that we can vary parametrically. Given a photographic image, we synthesized distinct images with identical model responses and asked whether human observers could discriminate them. From these data, we estimated the scaling constant that yields metameric images and found that it was consistent with receptive field sizes in area V2, suggesting a functional account of representation in that area.

Our model also provides an explanation for the phenomenon of visual crowding[18,19], in which humans fail to recognize peripherally presented objects surrounded by clutter. Crowding has been hypothesized to arise from compulsory pooling of peripheral information[20–23], and the development of our model was partly inspired by evidence that crowding is consistent with a representation based on local texture statistics[24]. Our model offers an instantiation of this hypothesis, providing a quantitative explanation for the spacing and eccentricity dependence of crowding effects, generalizing them to arbitrary photographic images and linking them to the underlying physiology of the ventral stream.

RESULTS

Our model is motivated by known facts about cortical computation, human pattern vision and the functional organization of ventral stream receptive fields. The V1 representation uses a bank of oriented filters covering the visual field, at all orientations and spatial frequencies. 'Simple' cells encode a single phase at each position, whereas 'complex' cells combine pairs of filters with the same preferred position, orientation and scale, but with different phase[25].

The second stage of our model achieves selectivity for compound image features by computing products between particular pairs of V1

[1]Center for Neural Science, New York University, New York, New York, USA. [2]Courant Institute of Mathematical Sciences, New York University, New York, New York, USA. [3]Howard Hughes Medical Institute, New York University, New York, New York, USA. Correspondence should be addressed to J.F. (freeman@cns.nyu.edu).

Received 9 May; accepted 22 June; published online 14 August 2011; doi:10.1038/nn.2889

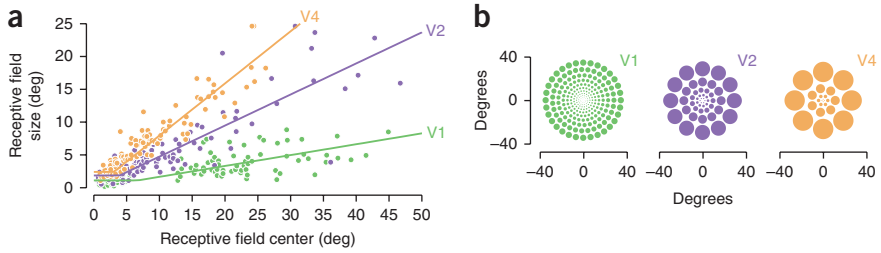


Figure 1 Physiological measurements of receptive field size in macaque. (**a**) Receptive field size (diameter) as a function of the distance between the receptive field center and the fovea (eccentricity) for visual areas V1, V2 and V4. Data were adapted from refs. 11 and 12, the only studies to measure receptive fields in all three macaque ventral stream areas with comparable methods. The size-to-eccentricity relationship in each area is well described by a 'hinged' line (see **Supplementary Analysis** for details and an analysis of a larger set of ten physiological datasets). (**b**) Cartoon depiction of receptive fields with sizes based on physiological measurements. The fovea is at the center of each array. The size of each circle is proportional to its eccentricity, based on the corresponding scaling parameter (slope of the fitted line in **a**). At a given eccentricity, a larger scaling parameter implies larger receptive fields. In our model, we used overlapping pooling regions (linear weighting functions) that uniformly tiled the image and were separable and of constant size when expressed in polar angle and log eccentricity (**Supplementary Fig. 1**).

responses (both simple and complex) and averaging these products over local regions, yielding local correlations. Correlations have been shown to capture important features of naturalistic texture images and have been used to explain some aspects of texture perception[17,26,27]. Correlations across orientations at different positions yield selectivity to angles and curved contours, as suggested by physiological studies of area V2 (refs. 28–32). Correlations across frequencies encode features with aligned phase or magnitude (for example, sharp edges or lines)[17,33], and correlations across positions capture periodicity. Finally, local correlations are compatible with models of cortical computation that propose hierarchical cascades of linear filtering, point nonlinearities and pooling[5–9,25,34,35] (see Online Methods).

We must specify the pooling regions over which pair-wise products of V1 responses are averaged. Receptive field sizes in the ventral stream grow approximately linearly with eccentricity, and the slope of this relationship (that is, the ratio of receptive field diameter to eccentricity) increases in successive areas (see **Fig. 1** and **Supplementary Analysis**). In our model, pooling is performed by weighted averaging, with smoothly overlapping functions that grow in size linearly with eccentricity, parameterized with a single scaling constant (see Online Methods and **Supplementary Fig. 1**).

Generation of metameric stimuli

If our model accurately describes the information captured (and discarded) at some stage of visual processing, and human observers cannot access the discarded information, then any two images that produce matching model responses should appear to be identical. To directly test this assertion, we examined perceptual discriminability of synthetic images that were as random as possible while producing identical model responses[17]. Model responses (**Fig. 2a**) were computed for a full-field photograph (for example, **Fig. 2b**). Synthetic images were then generated by initializing them with samples of Gaussian white noise and iteratively adjusting them (using a variant of gradient descent) until they matched the model responses of the original image (see Online Methods).

Synthetic images were identical to the original near the intended fixation point, where pooling regions were small, but features in the periphery were scrambled, and objects were grossly distorted and generally unrecognizable (**Fig. 2c**,**d**). When generated with the correct scaling constant, and viewed with proper fixation, however, the two images appeared to be nearly identical to the original and to each other.

Perceptual determination of critical scaling

To test the model more formally and to link it to a specific ventral stream area, we measured the perceptual discriminability of synthetic images as a function of the scaling constant used in their generation. If the model, with a particular choice of scaling constant, captures the information represented in some visual area, then model-generated stimuli will appear to be metameric. If the scaling constant is made larger, then the model will discard more information than the associated visual area and model-generated images will be readily distinguishable. If the model scaling is made smaller, then the model discards less information and the images will remain metameric. Thus, we sought the largest value of the scaling constant at which the stimuli appeared to be metameric. This critical scaling should correspond to the scaling of receptive field sizes in the area in which the information is lost.

As a separate control for the validity of this procedure, we examined stimuli generated from a V1 model that computes pooled V1 complex-cell responses[36] (that is, local spectral energy, see **Supplementary Fig. 2**). The critical scaling estimated for these stimuli should match the receptive field sizes of area V1. As the mid-ventral model includes a larger and more complex set of responses than the V1 model, we know a priori that the critical scaling for the mid-ventral model will be as large, or larger, than for the V1 model, but we do not know by how much.

For each model, we measured the ability of human observers to distinguish between synthetic images generated for a range of scaling constants (**Fig. 2e** and Online Methods). All four observers exhibited monotonically increasing performance as a function of scaling constant (**Fig. 3**). Chance performance (50%) indicates that the stimuli are metameric and, roughly speaking, the critical scaling is the value at which each curve first rises above chance.

To obtain an objective estimate of the critical scaling values, we derived an observer model that used the same ventral stream representation as was used to generate the matched images. The inputs to the observer model were two images that were matched over region sizes specified by scaling s. If we assume that the observer computes responses to each of these images with receptive fields that grow in size according to a fixed (but unknown) critical scaling s_0, then their ability to discriminate the two images will depend on the difference between the two sets of responses. We derived a closed-form expression for the dependency of this difference on s (see Online Methods). This expression is a function of the observer's scaling parameter, s_0, as well as a gain parameter, α_0, which controls their overall performance. We used signal detection theory[37] to describe the probability of a correct answer and fit the parameters (s_0, α_0) to the data of each subject by maximizing their likelihood.

The observer model provided an excellent fit to individual observer data for both the V1 and mid-ventral experiments (**Fig. 3**). Critical scaling values (s_0) were highly consistent across observers, with most

© 2011 Nature America, Inc. All rights reserved.

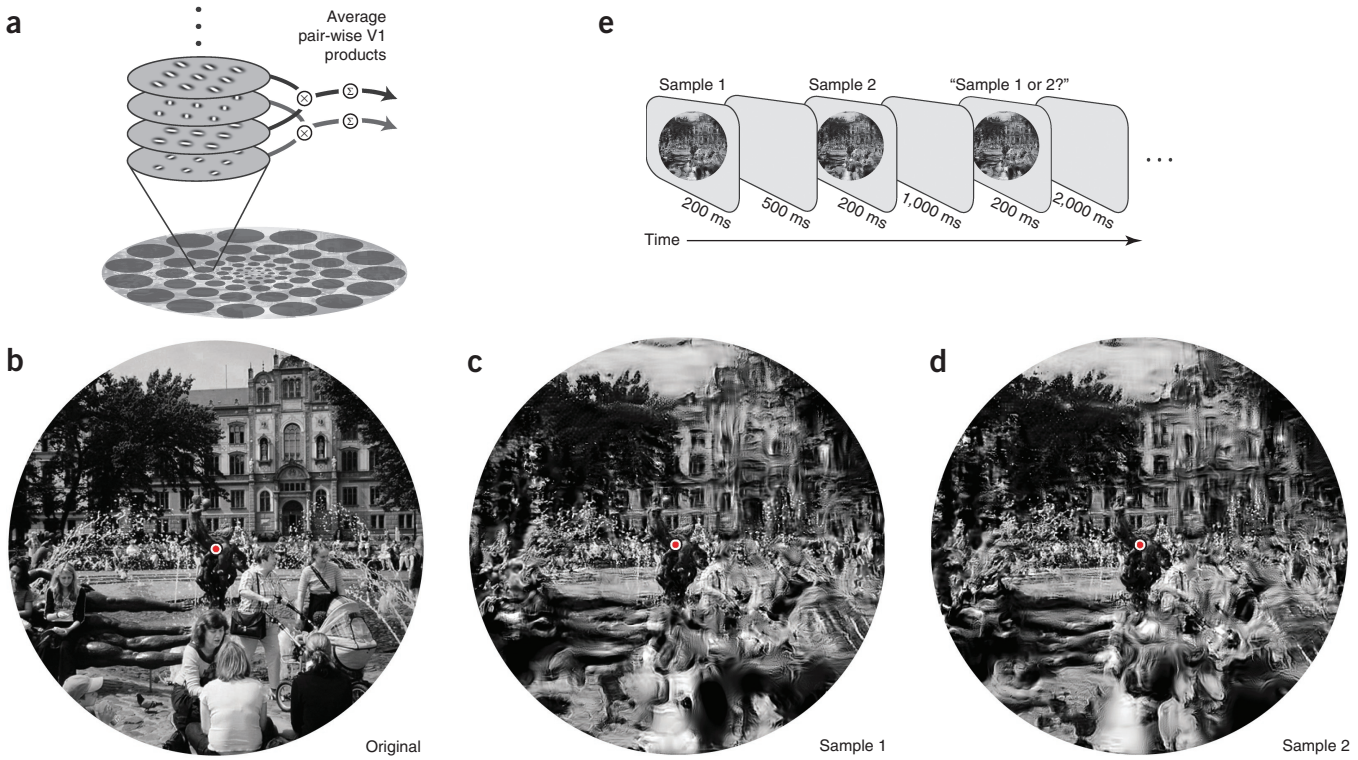
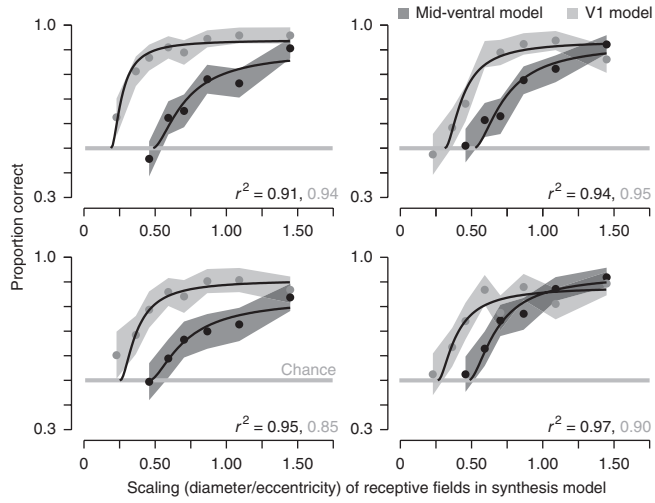


Figure 2 Mid-ventral model, example metameric stimuli and experimental task. (**a**) In each spatial pooling region, the image was first decomposed using a population of model V1 cells (both simple and complex), varying in their preferred orientation and spatial frequency. Model responses were computed from products of the filter outputs across different positions, orientations and scales, averaged over each of the pooling regions. (**b**) An original photograph of the Brunnen der Lebensfreude in Rostock, Germany. (**c**,**d**) Synthetic image samples, randomly selected from the set of images that generated model responses identical to those of the original (**b**). The value of the scaling parameter (used to determine the pooling regions of the model) was selected to yield 75% correct performance in discriminating such synthetic images (see **Fig. 4**). The two images, when viewed with fixation at the center (red dot), should appear to be nearly identical to the original and to each other, despite gross distortions in the periphery (for example, a woman's face is scrambled and dissolves into the spray of the fountain). (**e**) Psychophysical ABX task. Human observers viewed a sequence of two synthetic stimuli ABX, each randomly selected from the set of all images having model responses matched to an original image, followed by a third image that was identical to one of the first two. Observers indicated which of the first two images matched the third.

© 2011 Nature America, Inc. All rights reserved.

of the between-subject variability being captured by differences in overall performance (α_0). As expected, the simpler V1 model required a smaller scaling to generate metameric images. Specifically, critical scaling values for the V1 model were 0.26 ± 0.05 (mean ± s.d.), whereas values for the mid-ventral model were roughly twice as large (0.48 ± 0.02).



Estimation of physiological locus

We then compared the psychophysically estimated scaling parameters to physiological estimates of receptive field size scaling in different cortical areas. Functional magnetic resonance imaging has been used to measure population receptive fields in humans by estimating the spatial extent of a stimulus that contributes to the hemodynamic response across different regions of the visual field[13]. Although these sizes grow with eccentricity, and across successive visual areas, they include additional factors, such as variability in receptive field position and non-neural hemodynamic effects, which may depend on both eccentricity and visual area. We chose to compare our results

Figure 3 Metamer experiment results. Each graphs shows, for an individual observer, the proportion of correct responses in the ABX task as a function of the scaling parameter (ratio of receptive field diameter to eccentricity) of the model used to generate the stimuli. Data were averaged over stimuli drawn from four naturalistic images. Dark gray indicates the mid-ventral model (see **Fig. 2**), whereas light gray indicates the V1 model (see **Supplementary Fig. 2**). Shaded region indicates the 68% confidence interval obtained using bootstrapping. The gray horizontal lines indicate chance performance. Black lines indicate performance of observer model with critical scaling and gain parameters chosen to maximize the likelihood of the data for each individual observer (see Online Methods). r^2 values for the fits are indicated at the bottom of each plot.

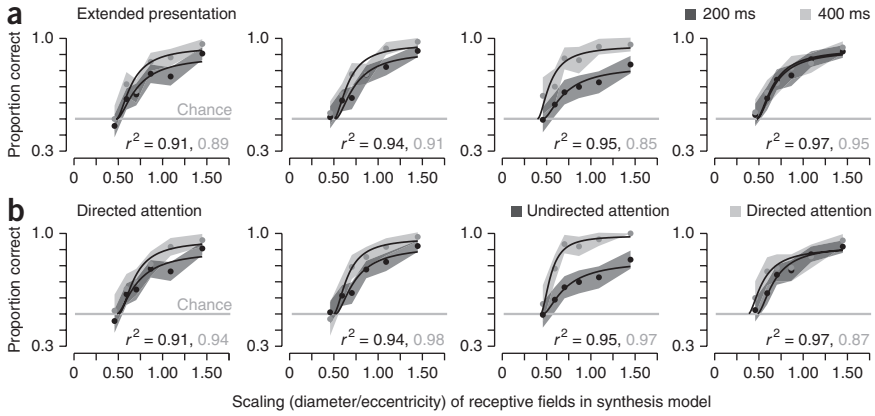


Figure 4 Metamer control experiments. Each column shows data and fitted psychometric functions for an individual observer. Both experiments used stimuli generated by the mid-ventral model. (**a**) Metamer experiment with extended presentation time. Light gray points indicate 400-ms presentation time, whereas dark gray points indicate 200-ms presentation time (replotted from **Fig. 3**). The shaded region represents the 68% confidence interval obtained using bootstrapping. The gray horizontal lines indicate chance performance. (**b**) Metamer experiment with directed attention. Light gray points indicate that observers were directed with an attentional cue indicating the region with the largest change (see Online Methods), whereas dark gray points represent undirected attention (replotted from **Fig. 3**). The shaded region represents the 68% confidence interval obtained using bootstrapping.

to single-unit electrophysiological measurements in non-human primates. Receptive field size estimates vary systematically, depending on the choice of stimuli and the method of estimation, so we combined estimates reported for ten different physiological datasets to obtain a distribution of scaling values for each visual area (see **Supplementary Analysis**). This analysis yielded values of 0.21 ± 0.07 for receptive fields in V1, 0.46 ± 0.05 for those of V2 and 0.84 ± 0.06 for those of visual area V4 (mean with 95% confidence intervals). Moreover, for studies that used comparable methods to estimate receptive fields in both V2 and V1, the average receptive field sizes in V2 were approximately twice the size of those in V1, for both macaques and humans[11,13,38].

As expected, the critical scaling value estimated from the V1 metamer experiment was well matched to the physiological estimates of receptive field scaling for V1 neurons. For the mid-ventral model, the critical scaling was roughly twice that of the V1 model, was well matched to receptive field sizes of V2 neurons and was substantially smaller than those of V4. This correspondence suggests that the metamerism of images synthesized using our mid-ventral model arises in area V2.

Robustness to bottom-up and top-down performance manipulations

If metamerism reflects a structural limitation of the visual system, governed by the eccentricity-dependent scaling of receptive field sizes, then the effects should be robust to experimental manipulations that alter observer performance without changing the spatial properties of the stimuli. To test this, we performed two variants of the mid-ventral metamer experiment, designed to alter performance through bottom-up and top-down manipulations of the experimental task.

First, we repeated the original experiment with doubled presentation times (400 ms instead of 200 ms). Fitting the observer model to data from four observers (**Fig. 4a**), we found that the gain parameter (α_0) was generally larger, accounting for increases in performance, but that the critical scaling (s_0) was statistically indistinguishable from that estimated in the original experiment ($P = 0.18$, two-tailed paired t test).

In a second control experiment, we manipulated endogenous attention. At the onset of each trial, a small arrow was presented at fixation, pointing toward the region in which the two subsequently presented stimuli differed most (see Online Methods). The fitted gain parameter was again generally larger, accounting for improvements in performance, but the critical scaling was statistically indistinguishable from that estimated in the original experiment ($P = 0.30$; two-tailed paired t test; **Fig. 4b**). In both control experiments, the increase in gain varied across observers and depended on their overall performance in the original experiment (some observers already had near-maximal performance). The scaling for the two control experiments were similar to those of the original experiment, were closely matched to the scaling of receptive fields found in area V2 and were much greater than the scaling found in the V1 metamer experiment ($P = 0.0064$, extended presentation task; $P = 0.0183$, attention task; two-tailed paired t test; **Fig. 5**).

Relationship to visual crowding

Our model predicts severe perceptual deficits in peripheral vision, some of which are revealed in the well-studied phenomenon of visual crowding[18,19], which has been hypothesized to arise from pooling or statistical combination in the periphery[20–24]. Crowding is typically characterized by asking observers to recognize a peripheral target object flanked by two distractors at varying target-to-flanker spacings. The critical spacing at which performance reaches threshold increases proportional to eccentricity[18,19], with reported rates ranging from 0.3 to 0.6. Our estimates of critical scaling for the mid-ventral model are in this range, but the substantial variability (which arises from different choices of stimuli, task, number of targets and flankers, and threshold) renders this comparison equivocal. Moreover, a direct comparison of these values may not even be warranted, as it implicitly relies on an unknown relationship between the pooling of the model responses and the degradation of recognition performance.

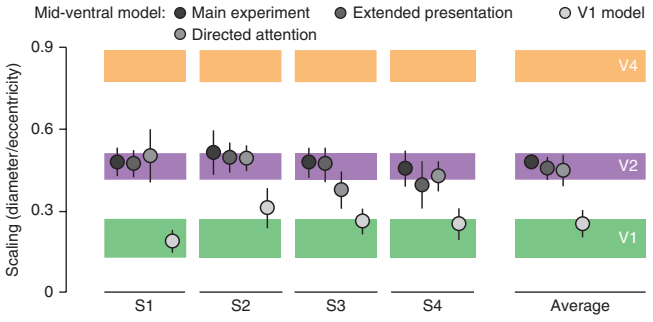


Figure 5 Summary of fitted critical scaling parameters for all experiments. Error bars indicate 95% confidence intervals on parameter estimates obtained through bootstrapping. Colored horizontal bars represent receptive field scaling as measured physiologically in each visual area, based on a meta-analysis combining across ten datasets (see **Supplementary Analysis** for details and references). The thickness of each bar indicates a 95% confidence interval.

© 2011 Nature America, Inc. All rights reserved.

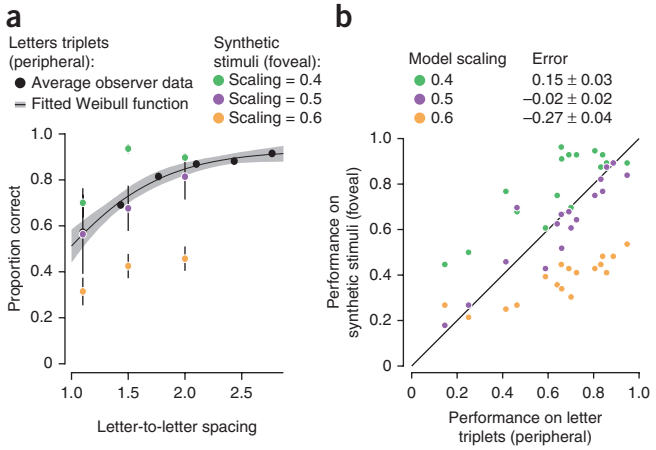


Figure 6 Crowding experiment. (**a**) Recognition performance for two different kinds of stimuli: peripherally viewed triplets of letters and foveally viewed stimuli synthesized to produce model responses identical to their corresponding letter triplets. Black dots represent the average recognition performance for a peripheral letter between two flankers, as a function of letter-to-letter spacing ($n = 5$ observers). The black line represents the best fitting Weibull function. The gray shaded region represents the 95% confidence interval for fit obtained through bootstrapping. Synthetic stimuli were generated for spacings yielding approximately 50%, 65% and 80% performance, based on the average psychometric function. Colored dots indicate average recognition performance for model-synthesized stimuli (foveally viewed). Different colors indicate the scaling parameter used in the model (purple, 0.5; orange, 0.6; green, 0.4). Error bars represent s.d. across observers. (**b**) Comparison of recognition performance for the peripheral letter triplets (from the psychometric function in **a**) and the foveally viewed synthetic stimuli (colored dots from **a**). Each point represents data from a single observer for a particular spacing and scaling. Two observers performed an additional condition at a larger eccentricity (not shown in **a**) to extend the range of performance levels (the six left-most points).

We performed an additional experiment to determine directly whether our mid-ventral model could predict recognition performance in a crowding task. The experimental design was inspired by a previous study linking statistical pooling in the periphery to crowding[24]. First, we measured observers' ability to recognize target letters presented peripherally (6 deg) between two flanking letters, varying the target-to-flanker spacing to obtain a psychometric function (**Fig. 6a**). We then used the mid-ventral model to generate synthetic metamers for a subset of these peripherally presented letter stimuli and measured the ability of observers to recognize the letters in these metamer stimuli under foveal viewing. Recognition failure (or success) for a single metamer cannot alone indicate crowding (or lack thereof), but the average performance across an ensemble of metamer samples quantifies the limitations on recognizability imposed by the model.

Average recognition performance for the metamers is well matched to that of their corresponding letter stimuli (**Fig. 6a**) for metamers synthesized with scaling parameter $s = 0.5$ (the average critical scaling estimated for our human observers). For metamers synthesized with scaling parameters of $s = 0.4$ or $s = 0.6$, performance was significantly higher or lower, respectively ($P < 0.0001$, two-tailed paired t test across observers and conditions). These results are consistent across all observers, at all spacings, and for two different eccentricities (**Fig. 6b**).

DISCUSSION

We constructed a model for visual pattern representation in the mid-level ventral stream that computes local correlations amongst V1 responses in eccentricity-dependent pooling regions. In addition, we developed a method for generating images with identical model responses and used these synthetic images to show that when the pooling region sizes of the model are set correctly, images with identical model responses are indistinguishable (metameric) to human observers, despite severe distortion of features in the periphery. We found that the critical pooling size required to produce metamericity is robust to bottom-up and top-down manipulations of discrimination performance; that critical pooling sizes are consistent with the eccentricity dependence of receptive field sizes of neurons in ventral visual area V2; and that the model can predict degradations of peripheral recognition known as crowding, as a function of both spacing and eccentricity.

Perceptual deficits in peripheral vision have been recognized for centuries. Most early studies focused on the loss of acuity that results from eccentricity-dependent sampling and blurring in the earliest visual stages. Crowding is a more complex peripheral deficit[39]. In 1976, Jerome Lettvin gave a subjective account of this phenomenon, describing letters embedded in text as having "lost form without losing crispness," and concluding that "the embedded [letter] only seems to have a 'statistical' existence."[20]. This article seems to have drifted into obscurity, but these ideas have been formalized in recent reports that explain crowding in terms of excessive averaging or pooling of features[21–24]. One study in particular hypothesized that crowding is a manifestation of the representation of peripheral visual content with local summary statistics[24], and showed that human recognition performance for crowded letters was matched to that of foveally viewed images synthesized to match the statistics of the original stimulus (computed over a localized region containing both the letter and flankers).

Our model provides an instantiation of these pooling hypotheses that operates over the entire visual field, which, in conjunction with the synthesis methodology, enabled several scientific advances. First, we validated the model with a metamer discrimination procedure, which provides a more direct test than comparisons to recognition performance in a crowding experiment. Second, the parameterization of eccentricity dependence allowed us to estimate the size of pooling regions and to associate the model with a distinct stage of ventral stream processing. Third, the full-field implementation allowed us to examine crowding in stimuli extending beyond a single pooling region and to account for the dependence of recognition on both eccentricity and spacing, the defining properties of crowding[18].

Finally, the fact that our model operates on arbitrary photographic images allows generalization of the laboratory phenomenon of crowding to complex scenes and everyday visual tasks. For example, crowding places limits on reading speed, as only a small number of letters around each fixation point are recognizable[40]. Model-synthesized metamers can be used to examine this 'uncrowded' window (**Fig. 7a**). We envision that our model could be used to optimize fonts, letter spacing or line spacing for robustness to crowding effects, potentially improving reading performance. There is also some evidence linking dyslexia to crowding with larger-than-normal critical spacing[18,41,42], and the model might serve as a useful tool for investigating this hypothesis. Model-synthesized images also show how camouflaged objects, which are already difficult to recognize foveally, blend into the background when viewed peripherally (**Fig. 7b,c**).

The interpretation of our experimental results relies on assumptions about the representation of, and access to, information in the brain. This is perhaps best understood by analogy to trichromacy[14]. Color metamers occur because information is lost by the cones and cannot be recovered in subsequent stages. However, color appearance judgments clearly do not imply direct conscious access to the responses of

© 2011 Nature America, Inc. All rights reserved.

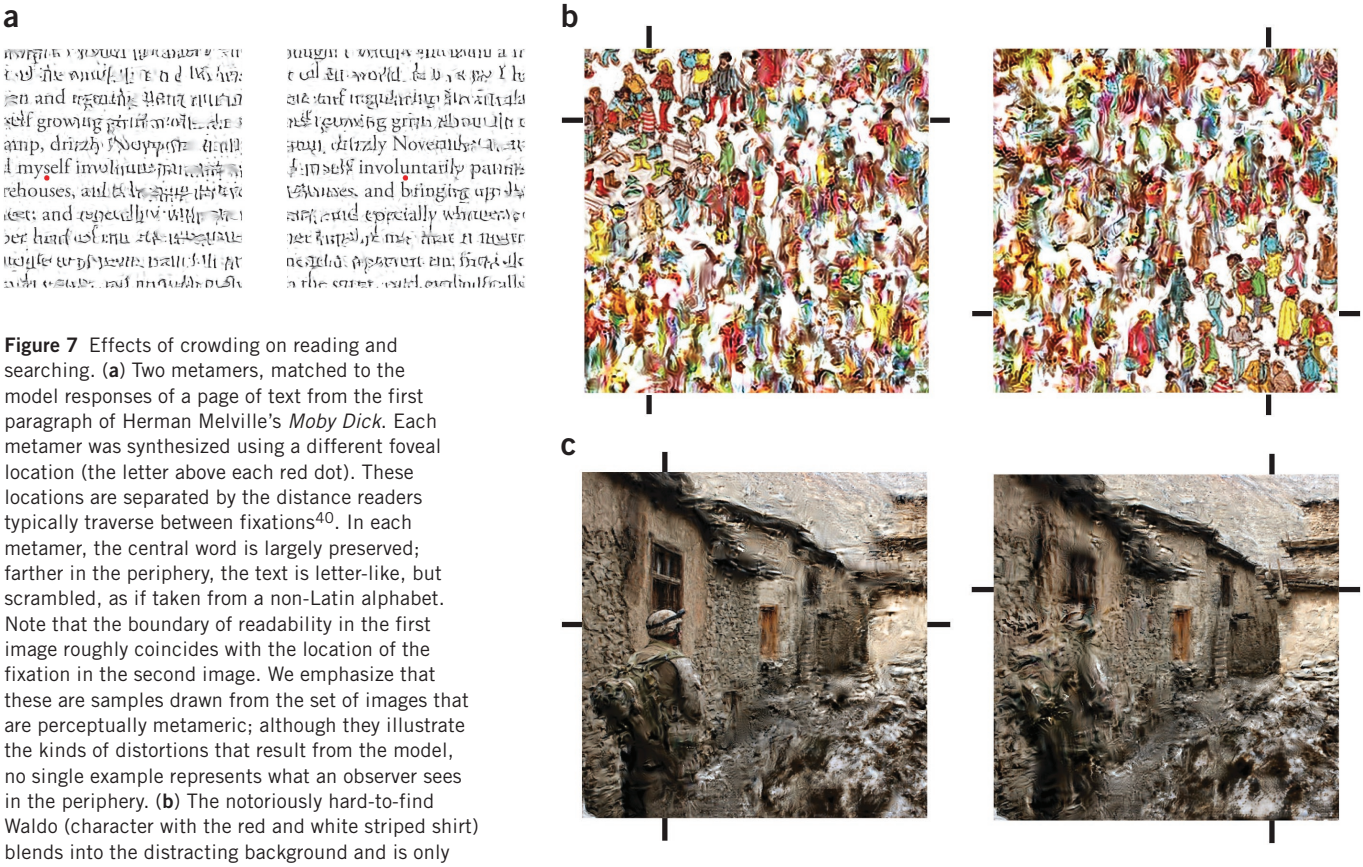


Figure 7 Effects of crowding on reading and searching. (**a**) Two metamers, matched to the model responses of a page of text from the first paragraph of Herman Melville's *Moby Dick*. Each metamer was synthesized using a different foveal location (the letter above each red dot). These locations are separated by the distance readers typically traverse between fixations[40]. In each metamer, the central word is largely preserved; farther in the periphery, the text is letter-like, but scrambled, as if taken from a non-Latin alphabet. Note that the boundary of readability in the first image roughly coincides with the location of the fixation in the second image. We emphasize that these are samples drawn from the set of images that are perceptually metameric; although they illustrate the kinds of distortions that result from the model, no single example represents what an observer sees in the periphery. (**b**) The notoriously hard-to-find Waldo (character with the red and white striped shirt) blends into the distracting background and is only recognizable when fixated. Cross-hairs surrounding each image indicate the location of the fovea used by the model during synthesis. (**c**) A soldier in Afghanistan wears patterned clothing to match the stoney texture of the environment and similarly blends into the background.

© 2011 Nature America, Inc. All rights reserved.

those cones. Analogously, our experiments imply that the information loss ascribed to areas V1 and V2 cannot be recovered or accessed by subsequent stages of processing (two stimuli that are V1 metamers, for example, should also be V2 metamers). However, this does not imply that observers directly access the information represented in V1 or V2. Indeed, if observers could access V1 responses, then any additional information loss incurred when those responses are combined and pooled in V2 would have no perceptual consequence and the stimuli generated by the mid-ventral model would not appear to be metameric.

The loss of information in our model arises directly from its architecture, the set of statistics and the pooling regions over which they are computed, and this determines the set of metameric stimuli. Discriminability of non-metameric stimuli depends on the strength of the information preserved by the model, relative to noise. As seen in the presentation time and attention control experiments, manipulations of signal strength did not alter the metamericity of stimuli and therefore did not affect estimates of critical scaling. These results are also consistent with the crowding literature. Crowding effects are robust to presentation time[43], and attention can increase performance in crowding tasks while yielding small or no changes in critical spacing[19,44]. Certain kinds of exogenous cues, however, may reduce critical spacing[45], and perceptual learning has been shown to reduce critical spacing through several days of intensive training[46]. If either manipulation were found to reduce critical scaling (as estimated from a metamer discrimination experiment), we would interpret this as arising from a reduction in receptive field sizes, which could be verified through electrophysiological measurements.

From a physiological perspective, our model is deliberately simplistic. We expect that incorporating more realistic response properties (for example, spike generation, feedback circuitry) would not substantially alter the information represented in model populations, but would render the synthesis of stimuli computationally intractable. Despite the simplicity of the model, the metamer experiments do not uniquely constrain the response properties of individual model neurons. This may again be understood by analogy with the case of trichromacy: color-matching experiments constrain the linear subspace spanned by the three cone absorption spectra, but do not uniquely constrain the spectra of the individual cones[14]. Thus, identification of V2 as the area in which the model resides does not imply that responses of individual V2 neurons encode local correlations. Our results, however, do suggest new forms of stimuli that could be used to explore such responses in physiological experiments. In a single pooling region, the model provides a parametric representation of local texture features[17]. Stochastic stimuli containing these features are more complex than sine gratings or white noise, but better controlled (and more hypothesis driven) than natural scenes or objects, and are therefore well suited for characterizing responses of individual cells[47].

Finally, one might ask why the ventral stream discards such a substantial amount of information. Theories of object recognition posit that the growth of receptive field sizes in consecutive areas, as well as with eccentricity, confers invariance to geometric transformations, and cascaded models based on filtering, simple nonlinearities and successively broader spatial pooling have been used to explain such invariances measured in inferotemporal cortex[8–10,48]. Our model closely resembles the early stages of these models, but

our inclusion of eccentricity-dependent pooling and the invariance to feature scrambling revealed by the metamericity of our synthetic stimuli seems to be at odds with the goal of object recognition. One potential resolution of this conundrum is that the two forms of invariance arise in distinct parallel pathways. An alternative possibility is that a texture-like representation in the early ventral stream provides a substrate for object representations in later stages. Such a notion was suggested by Lettvin, who hypothesized that "texture, somewhat redefined, is the primitive stuff out of which form is constructed."[20]. If so, the metamer procedure introduced here provides a powerful tool for exploring the nature of invariances arising in subsequent stages of the ventral stream.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We would like to thank R. Rosenholtz for early inspiration and discussions regarding the relationship between texture and crowding, N. Rust for discussions about the nature of information represented in the ventral stream, C. Anderson for discussions about the scaling of receptive fields with eccentricity, M. Landy, A. Girshick and R. Goris for advice on experimental design, C. Ekanadham and U. Rajashaker for advice on the model and analysis, and D. Ganguli, D. Heeger, J. McDermott, E. Merriam and C. Ziemba for comments on the initial manuscript. This work was supported by a National Science Foundation Graduate Student Fellowship to J.F. and a Howard Hughes Medical Institute Investigatorship to E.P.S.

AUTHOR CONTRIBUTIONS

J.F. and E.P.S. conceived the project and designed the experiments. J.F. implemented the model, performed the experiments and analyzed the data. J.F. and E.P.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

© 2011 Nature America, Inc. All rights reserved.

1. Ungerleider, L.G. & Haxby, J.V. 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* **4**, 157–165 (1994).
2. Hubel, D.H. Exploration of the primary visual cortex, 1955–78. *Nature* **299**, 515–524 (1982).
3. Carandini, M. *et al.* Do we know what the early visual system does? *J. Neurosci.* **25**, 10577–10597 (2005).
4. Tanaka, K. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* **19**, 109–139 (1996).
5. Granlund, G. In search of a general picture processing operator. *Comput. Graph. Image Process.* **8**, 155–173 (1978).
6. Fukushima, K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
7. LeCun, Y. *et al.* Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **2**, 396–404 (1989).
8. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
9. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. & Poggio, T. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 411–426 (2007).
10. Rolls, E. The neurophysiology and computational mechanisms of object representation. in *Object Categorization: Computer and Human Vision Perspectives* (eds. Dickinson, S.J., Leonardis, A., Schiele, B. & Tarr, M.J.) 257–287 (Cambridge University Press, 2009).
11. Gattass, R., Gross, C.G. & Sandell, J.H. Visual topography of V2 in the macaque. *J. Comp. Neurol.* **201**, 519–539 (1981).
12. Gattass, R., Sousa, A.P. & Gross, C.G. Visuotopic organization and extent of V3 and V4 of the macaque. *J. Neurosci.* **8**, 1831–1845 (1988).
13. Dumoulin, S.O. & Wandell, B.A. Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647–660 (2008).
14. Wandell, B. *Foundations of Vision* (Sinauer Associates, 1995).
15. Julesz, B. Visual pattern discrimination. *IEEE Trans. Inf. Theory* **8**, 84–92 (1962).
16. Koenderink, J. & Doom, A.J.V. Local image operators and iconic structure. *Algebr. Frames Percept. Action Cycle* **1315**, 66–93 (1997).
17. Portilla, J. & Simoncelli, E.P. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).
18. Pelli, D.G. & Tillman, K.A. The uncrowded window of object recognition. *Nat. Neurosci.* **11**, 1129–1135 (2008).
19. Levi, D.M. Crowding–an essential bottleneck for object recognition: a mini-review. *Vision Res.* **48**, 635–654 (2008).
20. Lettvin, J.Y. On seeing sidelong. *The Sciences* **16**, 10–20 (1976).
21. Parkes, L., Lund, J., Angelucci, A., Solomon, J.A. & Morgan, M. Compulsory averaging of crowded orientation signals in human vision. *Nat. Neurosci.* **4**, 739–744 (2001).
22. Pelli, D.G., Palomares, M. & Majaj, N.J. Crowding is unlike ordinary masking: distinguishing feature integration from detection. *J. Vis.* **4**, 1136–1169 (2004).
23. Greenwood, J.A., Bex, P.J. & Dakin, S.C. Positional averaging explains crowding with letter-like stimuli. *Proc. Natl. Acad. Sci. USA* **106**, 13130–13135 (2009).
24. Balas, B., Nakano, L. & Rosenholtz, R. A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* **9**, 13 (2009).
25. Adelson, E.H. & Bergen, J.R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284–299 (1985).
26. Graham, N. *Visual Pattern Analyzers* (Oxford University Press, 1989).
27. Balas, B. Attentive texture similarity as a categorization task: comparing texture synthesis models. *Pattern Recognit.* **41**, 972–982 (2008).
28. Hegdé, J. & Essen, D.C.V. Selectivity for complex shapes in primate visual area V2. *J. Neurosci.* **20**, RC61 (2000).
29. Ito, M. & Komatsu, H. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neurosci.* **24**, 3313–3324 (2004).
30. Anzai, A., Peng, X. & Essen, D.C.V. Neurons in monkey visual area V2 encode combinations of orientations. *Nat. Neurosci.* **10**, 1313–1321 (2007).
31. Schmid, A.M., Purpura, K.P., Ohiorhenuan, I.E., Mechler, F. & Victor, J.D. Subpopulations of neurons in visual area v2 perform differentiation and integration operations in space and time. *Front. Syst. Neurosci.* **3**, 15 (2009).
32. Willmore, B.D.B., Prenger, R.J. & Gallant, J.L. Neural representation of natural images in visual area V2. *J. Neurosci.* **30**, 2102–2114 (2010).
33. Kovesi, P. Phase congruency: a low-level image invariant. *Psychol. Res.* **64**, 136–148 (2000).
34. Simoncelli, E.P. & Heeger, D.J. A model of neuronal responses in visual area MT. *Vision Res.* **38**, 743–761 (1998).
35. David, S.V., Hayden, B.Y. & Gallant, J.L. Spectral receptive field properties explain shape selectivity in area V4. *J. Neurophysiol.* **96**, 3492–3505 (2006).
36. Chen, X., Han, F., Poo, M.-M. & Dan, Y. Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). *Proc. Natl. Acad. Sci. USA* **104**, 19120–19125 (2007).
37. Macmillan, N.A., Kaplan, H.L. & Creelman, C.D. The psychophysics of categorical perception. *Psychol. Rev.* **84**, 452–471 (1977).
38. Shushruth, S., Ichida, J.M., Levitt, J.B. & Angelucci, A. Comparison of spatial summation properties of neurons in macaque V1 and V2. *J. Neurophysiol.* **102**, 2069–2083 (2009).
39. Bouma, H. Interaction effects in parafoveal letter recognition. *Nature* **226**, 177–178 (1970).
40. Pelli, D.G. *et al.* Crowding and eccentricity determine reading rate. *J. Vis.* **7**, 20 (2007).
41. Geiger, G., Lettvin, J.Y. & Zegarra-Moran, O. Task-determined strategies of visual process. *Brain Res. Cogn. Brain Res.* **1**, 39–52 (1992).
42. Martelli, M., Filippo, G.D., Spinelli, D. & Zoccolotti, P. Crowding, reading, and developmental dyslexia. *J. Vis.* **9**, 1–14 (2009).
43. Townsend, J.T., Taylor, S.G. & Brown, D.R. Lateral masking for letters with unlimited viewing time. *Atten. Percept. Psychophys.* **10**, 375–378 (1971).
44. Scolari, M., Kohnen, A., Barton, B. & Awh, E. Spatial attention, preview, and popout: which factors influence critical spacing in crowded displays? *J. Vis.* **7**, 7 (2007).
45. Yeshurun, Y. & Rashal, E. Precueing attention to the target location diminishes crowding and reduces the critical distance. *J. Vis.* **10**, 16 (2010).
46. Chung, S.T.L. Learning to identify crowded letters: does it improve reading speed? *Vision Res.* **47**, 3150–3159 (2007).
47. Rust, N.C. & Movshon, J.A. In praise of artifice. *Nat. Neurosci.* **8**, 1647–1650 (2005).
48. Zoccolan, D., Kouh, M., Poggio, T. & DiCarlo, J.J. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* **27**, 12292–12307 (2007).

ONLINE METHODS

Multi-scale multi-orientation decomposition. Images were partitioned into subbands by convolving with a bank of filters tuned to different orientations and spatial frequencies. We used the 'steerable pyramid', which has several advantages over common alternatives (for example, Gabor filters and orthogonal wavelets), including direct reconstruction properties (beneficial for synthesis), translation invariance in subbands and rotation invariance across orientation bands[17]. A MATLAB (MathWorks) implementation is available at http://www.cns.nyu.edu/~lcv/software.php. The filters were directional third derivatives of a low-pass kernel and were spatially localized, oriented, antisymmetric and roughly one octave in spatial frequency bandwidth. We used a set of 16 filters, rotated and dilated to cover four orientations and four scales. We also included a set of even-symmetric filters of identical Fourier amplitude (that is, Hilbert transforms)[17]. Each subband was subsampled at its associated Nyquist frequency, so that filter spacing was proportional to size. We write the nth subband as $x_n(i, j)$, a two-dimensional array containing the complex-valued responses. We also use vector notation $\vec{\mathbf{x}}_n$. The real part of the subband is denoted as $s_n(i, j)$ and represents the responses of V1 simple cells. The square root of the sum of the squared responses of symmetric and antisymmetric filters yields a phase-invariant measure of local magnitude, denoted $e_n(i, j)$, and represents the responses of V1 complex cells[17,25].

Mid-ventral model. The second stage of the model computes products of pairs of V1 responses tuned to neighboring orientations, scales and positions. The model responses were based on those developed previously[17] for global texture modeling, but all averages were computed over localized pooling regions. A pooling region is defined by a positive-valued weighting function denoted $w(i, j)$, whose values sum to 1 (see below for details on their construction). The mid-ventral model includes, first, the products of responses at nearby spatial locations (that is, autocorrelations) for both simple cells (capturing spectral features such as periodicity) and complex cells (capturing spatially displaced occurrences of similarly oriented features). Simple cell autocorrelations are given by

$$A_w(n,k,l) = \sum \sqrt{w(i,j)}\left(s_n(i,j) - \mu_w(\vec{\mathbf{s}}_n)\right) \times \sqrt{w(i+k,j+l)}\left(s_n(i+k,j+l) - \mu_w(\vec{\mathbf{s}}_n)\right) \quad (1)$$

where (k,l) specifies the spatial displacement (in horizontal and vertical directions), the summation is over (i, j), and $\mu_w(\vec{\mathbf{s}}_n)$ is the weighted mean

$$\mu_w(\vec{\mathbf{s}}_n) = \sum w(i,j) s_n(i,j) \quad (2)$$

Complex cell autocorrelations are similarly given by

$$B_w(n,k,l) = \sum \sqrt{w(i,j)}\left(e_n(i,j) - \mu_w(\vec{\mathbf{e}}_n)\right) \times \sqrt{w(i+k,j+l)}\left(e_n(i+k,j+l) - \mu_w(\vec{\mathbf{e}}_n)\right) \quad (3)$$

In the mid-ventral model, we included spatial displacements in the range $(-3 \le k \le 3,\ -3 \le l \le 3)$ for both autocorrelations. In the V1 model (see below), we only included the central sample (that is, $k = l = 0$) for which equations (1) and (2) reduce to weighted variances.

Second, the model includes products of complex cell responses with those at other orientations (capturing structures with mixed orientation content, such as junctions or corners) and with those at adjacent scales (capturing oriented features with spatially sharp transitions such as edges, lines and contours). These cross-correlations are given by

$$C_w(n,m) = \sum w(i,j)\left(e_n(i,j) - \mu_w(\vec{\mathbf{e}}_n)\right)\left(e_m(i,j) - \mu_w(\vec{\mathbf{e}}_n)\right) \quad (4)$$

where indices (n, m) specify two subbands arising from filters at different orientations at the same scale or at different orientations at adjacent scales. This yields six cross-orientation correlations at each scale and 16 cross-scale correlations for each scale.

Third, the model includes products of the simple cell responses with phase-doubled simple cell responses at the next coarsest scale. Phase relationships at adjacent scales distinguish lines from edges and can also capture gradients in intensity arising from shading[17]. These correlations are given by

$$S_w(n,m) = \sum w(i,j)\left(x_n(i,j) - \mu_w(\vec{\mathbf{x}}_n)\right)\left(\frac{x_m^2(i,j)}{|x_m(i,j)|} - \mu_w\left(\frac{x_m^2(i,j)}{|x_m(i,j)|}\right)\right) \quad (5)$$

where indices (n, m) specify two adjacent scales (n is the finer scale).

It is worth noting that all of these products may be represented equivalently as differences of squared sums and differences (that is, $4ab=(a+b)^2-(a-b)^2$), which might provide a more physiologically plausible form[25]. We also included three weighted marginal statistics (variance, skew and kurtosis) of the low-pass images reconstructed at each scale of the course-to-fine process. The weighted mean is given in equation (2). Higher-order weighted moments of order p are

$$\mu_w^p(\vec{\mathbf{s}}_n) = \sum w(i,j)\left(s_n(i,j) - \mu_w(\vec{\mathbf{s}}_n)\right)^p \quad (6)$$

From this, the skew and kurtosis are

$$\gamma_w(\vec{\mathbf{s}}_n) = \frac{\mu_w^3(\vec{\mathbf{s}}_n)}{(\mu_w^2(\vec{\mathbf{s}}_n))^{3/2}} \quad (7)$$

$$\kappa_w(\vec{\mathbf{s}}_n) = \frac{\mu_w^4(\vec{\mathbf{s}}_n)}{(\mu_w^2(\vec{\mathbf{s}}_n))^2} \quad (8)$$

Pooling regions. Each of the model statistics was computed using locally weighted spatial averages. The weighting functions (generically denoted $w(i, j)$ in the preceding section) are smooth and overlapping, and arranged so as to tile the image (that is, they sum to a constant). These functions are separable with respect to polar angle and log eccentricity, ensuring that they grow linearly in size with eccentricity (see examples in **Supplementary Fig. 1**). Weighting in each direction is defined in terms of a generic 'mother' window, with a flat top and squared cosine edges.

$$f(x) = \begin{cases} \cos^2\left(\frac{\pi}{2}\left(\frac{x-(t-1)/2}{t}\right)\right), & -(1+t)/2 < x \le (t-1)/2 \\ 1, & (t-1)/2 < x \le (1-t)/2 \\ -\cos^2\left(\frac{\pi}{2}\left(\frac{x-(1+t)/2}{t}\right)\right)+1, & (1-t)/2 < x \le (1+t)/2 \end{cases} \quad (9)$$

These window functions sum to a constant when spaced on the unit lattice. The parameter t specifies transition region width, and is set to 1/2 for our experiments. For polar angles, we require an integer number N_θ of windows between 0 and π. The full set is

$$h_n(\theta) = f\left(\frac{\theta - \left(w_\theta n + \frac{w_\theta(1-t)}{2}\right)}{w_\theta}\right), \quad w_\theta = \frac{2\pi}{N_\theta}, \quad n = 0 \ldots N_\theta - 1 \quad (10)$$

where n indexes the windows and w_θ is width. For log eccentricity, an integer number of windows is not required. However, to equate boundary conditions across scaling conditions in our experiments, we centered the outermost window on the radius of the image (e_r). For computational efficiency, we also did not include windows below a minimum eccentricity (e_0 – approximately half a degree of visual angle in our stimuli). For smaller eccentricities, pooling regions are extremely small and constrain the model to reproduce the original image. Between the minimum and maximum eccentricities, we constructed N_e windows

$$g_n(e) = f\left(\frac{\log(e) - [\log(e_0) + w_e(n+1)]}{w_e}\right), \quad w_e = \frac{\log(e_r) - \log(e_0)}{N_e}, \quad n = 0 \ldots N_e - 1 \quad (11)$$

where n indexes the windows and w_e is the width. The number of windows, N_e, determines the ratio of radial full-width at half-maximum to eccentricity, which

© 2011 Nature America, Inc. All rights reserved.

doi:10.1038/nn.2889

is reported as the scaling (for example, **Figs. 4 and 5**). We can achieve an arbitrary scaling (that is, a non-integer number of windows) by releasing the constraint on the endpoint location (for example, **Figs. 6 and 7**). For each choice of scaling, we chose an integer number of polar-angle windows (N_θ) that yielded an aspect ratio of radial width to circumferential width of approximately 2. There are few studies on peripheral receptive field shape in the ventral stream, but our choice was motivated by reports of radially elongated receptive fields and radial biases throughout the visual system[49,50]. Future work could explore the effects of both the scaling and the aspect ratio on metamericity.

To use each window at different scales of the pyramid, we create an original window in the pixel domain and then generate low-pass windows to be applied at different scales by blurring and sampling the original (that is, we construct a Gaussian pyramid). The information captured by averages computed with this full set of two-dimensional windows is approximately invariant to global rotation or dilation: shifting the origin of the log-polar coordinate system in which they are defined would re-parameterize the model without substantially changing the class of metameric stimuli corresponding to a particular original image.

V1 model. The model for our V1 control experiment uses the same components described above. We used the same linear filter decomposition, and then squared and pooled these responses directly, consistent with physiological characterizations in V1 (ref. 36). This model does not include the local correlations (that is, pair-wise products) used in the mid-ventral model. Both the V1 model and the mid-ventral model collapse the computation into a single stage of pooling, instead of building the mid-ventral model on the responses of a pooled V1 stage (and previous stages, such as the retina and LGN). This kind of simplification is common in modeling sensory representations and allowed us to develop a tractable synthesis procedure.

Synthesis. Metameric images were synthesized to match a set of measurements made on an original image. An image of Gaussian white noise was iteratively adjusted until it matched the model responses of the original. Synthesizing from different white noise samples yields distinct images. This procedure approximates sampling from the maximum entropy distribution over images matched to a set of model responses[17]. We used gradient descent to perform the iterative image adjustments. For each set of responses, we computed gradients, following previous derivations[17], but including the effects of the window functions. Descent steps were taken in the direction of these gradients, starting with the low-frequency subbands (that is, coarse to fine). For autocorrelations, gradients for each pooling region were combined to give a global image gradient on each step. Gradient step sizes were chosen to stabilize convergence. For the cross correlations, single-step gradient projections were applied to each pooling region iteratively.

We used 50 iterations for all of the images generated for the experiments. Parameter convergence was verified by measuring mean squared error, normalized by the parameter variance. For samples synthesized from the same original image, this metric was 0.01 ± 0.015 (mean ± s.d.) across all images and scalings used in our experiments. As an indication of computational cost, synthesis of a 512×512 pixel image for a scaling of $s = 0.5$ took approximately 6 to 8 h on a Linux workstation with 2.6 GHz dual Opteron 64-bit processor and 32 GB RAM. Smaller scaling values require more windows, and thus more parameters and more time. The entire set of experimental stimuli took approximately 1 month of compute time to generate. Synthesis sometimes required more steps to converge for artificial stimuli, such as those created for the crowding experiments (**Fig. 6**), so we used 100 iterations for those syntheses.

Experimental stimuli. Stimuli were derived from four naturalistic photographs from the authors' personal collection. One image depicts a natural scene (trees and shrubbery) and the other three depict people and man-made objects. For each photograph, we synthesized three images for each of six values of the scaling parameter s. Pilot data revealed that performance was at chance for the smallest value tested, so we did not generate stimuli at smaller scalings. The V1 model was simpler, allowing us to synthesize stimuli for three smaller scaling values.

Psychophysics. Eight observers (ages 24–32, six males, two females) with normal or corrected-to-normal vision participated in the experiment. Protocols for selection of observers and experimental procedures were approved by the human subjects committee of New York University and all subjects signed an approved consent form. One observer was an author; all others were naive to the purposes of the experiment. Four observers participated in the metamer experiments (described in this section), and five observers participated in the crowding experiments (described below). One observer participated in both.

In the metamer experiments, two observers (S3 and S4) were tested with eye tracking (see below), with stimuli presented on a 22-inch flat screen CRT monitor at a distance of 57 cm. Two observers (S1 and S2) were tested without eye tracking, with stimuli presented on a 13-inch flat screen LCD monitor at a distance of 38 cm. In both displays, all images were presented in a circular window subtending 26 degrees of visual angle and blended into the background with a 0.75-degree-wide raised cosine. A 0.25-degree fixation square was shown throughout the experiment.

Each trial of the ABX task (**Fig. 3**) used two different synthesized image samples, matched to the model responses of a corresponding original image. At the start of each trial, the observer saw one image for 200 ms. After a 500-ms delay, the observer saw the second image for 200 ms. After a 1,000-ms delay, the observer saw one of the two images repeated, for 200 ms. The observer indicated with a key press whether the third image looked more like the first (1) or the second (2). There was no feedback during the experiment. Before the experiment, each observer performed a small number of practice trials (~5) with feedback to become familiar with the task.

In the mid-level ventral experiment, we used four original images and six scaling conditions, and created three synthetic images for each original/scaling combination. This yielded 12 unique ABX sequences per condition. In each block of the experiment, observers performed 288 trials, one for each combination of image (4), scaling (6) and trial type (12). Observers performed four blocks (1,152 trials). The V1 experiment was identical, except that it included nine scaling conditions, resulting in 384 trials per block. Observers performed three blocks (1,152 trials). Blocks were performed on different days, so the observer never saw the same stimulus sequence twice in the same session. Psychometric functions and parameter estimates were similar across blocks, suggesting that observers did not learn any particular image feature. Results were also similar across the four original images and were thus combined.

We performed two further control experiments using the stimuli from the mid-ventral metamer experiment. The first of these was identical to the main experiment except that presentation time was lengthened to 400 ms. Each observer performed either two or three blocks (576 or 864 trials). The second experiment was identical to the main experiment, except that at the beginning of each trial a small line (1 deg long) emanating from fixation was presented for 300 ms, with a 300-ms blank period before and after. On each trial, we computed the squared error (in the pixel domain) between the two to-be-presented images and averaged the squared error in each of six radial sections. The line cue pointed to the section with largest squared error. Each observer performed two blocks (576 trials).

Eye tracking. Two observers (S3 and S4) were tested while their gaze positions were measured (500 Hz, monocular) with an Eyelink 1000 (SR Research) eye tracker, for all four metamer experiments. A 9-point calibration was performed at the start of each block. We analyzed the eye position data to discard trials with broken fixation. We first computed a fixation location for each block by averaging eye positions over all trials. This was used as fixation, rather than the physical screen center, to account for systematic offset due to calibration error. We then computed, on each trial, the distance of each gaze position from fixation; a trial was discarded if this distance exceeded 2 degrees for any gaze position. We discarded 5% (S3) and 17% (S4) of trials across all four experiments. Using a more conservative (1 degree) threshold discarded more trials, but did not substantially change psychometric functions or critical scaling estimates. By only including trials with stable fixation, we ruled out the possibility that systematic differences in fixation among scaling conditions, presentation conditions or models could account for our results.

Fitting the psychometric function. We assumed that an observer's performance in the ABX experiment was determined by a population of mid-ventral neurons whose receptive fields grew with eccentricity according to scaling parameter s_0, and their performance depended on the total squared difference of those responses computed on the two presented images generated with model critical scaling s. We derived a closed-form approximation to that squared difference as a function of s_0 and s. Let $\vec{\mathbf{x}}$ be a vector of values from an original image to be locally

© 2011 Nature America, Inc. All rights reserved.

averaged (for example, a vector containing pair-wise products of two orientation subbands). Let M be a matrix whose rows contain the weighting functions (with sizes scaling according to s) that are used to compute local averages. Assume that a second vector $\vec{\mathbf{y}}$ was initially set to a vector of white noise samples, $\vec{\mathbf{n}}$, and then adjusted so that $M\vec{\mathbf{x}} = M\vec{\mathbf{y}}$, that is, the two images match with respect to the local averages computed by M. Define the projection matrix $P=M^T(MM^T)^{-1}M$, which projects vectors into the space spanned by M. We can rewrite $\vec{\mathbf{y}}$ as the sum of two components

$$\vec{\mathbf{y}} = (I-P)\vec{\mathbf{n}} + P\vec{\mathbf{x}} \tag{12}$$

where the first term is the component of $\vec{\mathbf{n}}$ that lies in the null space of M, and the second is constrained by the fact that $\vec{\mathbf{y}}$ is matched to $\vec{\mathbf{x}}$ (that is, $M\vec{\mathbf{x}} = M\vec{\mathbf{y}}$).

Now let R be the matrix that the observer uses to compute averages over regions scaling with s_0. We assumed that the discriminability of the two stimuli depends on the sum of squared differences between these averages. We can express the expected value of this quantity, taken over instantiations of $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ that match the same model measurements, as

$$\begin{aligned} d^2 &= \mathbf{E}\left[\left\|R\vec{\mathbf{x}} - R\vec{\mathbf{y}}\right\|^2\right] \\ &= \mathbf{E}\left[\left\|R\left(\left((I-P)\vec{\mathbf{x}} + P\vec{\mathbf{x}}\right) - \left((I-P)\vec{\mathbf{n}} + P\vec{\mathbf{x}}\right)\right)\right\|^2\right] \\ &= \mathbf{E}\left[\left\|R(I-P)(\vec{\mathbf{x}} - \vec{\mathbf{n}})\right\|^2\right] \end{aligned} \tag{13}$$

where we use the definition of $\vec{\mathbf{y}}$ from equation (12) and rewrite $\vec{\mathbf{x}}$ in a similar form. Assuming that $\vec{\mathbf{x}}$ and $\vec{\mathbf{n}}$ are independent and have the same covariance matrix C, we obtain

$$\begin{aligned} d^2 &= Tr\left(\mathbf{E}\left[R(I-P)(\vec{\mathbf{x}} - \vec{\mathbf{n}})(\vec{\mathbf{x}} - \vec{\mathbf{n}})^T (I-P)^T R^T\right]\right) \\ &= Tr\left(R(I-P)2C(I-P^T)R^T\right) \\ &= Tr\left(\left(R - RM^T(MM^T)^{-1}M\right)2C\left(R^T - M^T(MM^T)^{-1}MR^T\right)\right) \end{aligned} \tag{14}$$

We can obtain a simple functional form for this expression by assuming that C is a multiple of the identity matrix. In general, the components of $\vec{\mathbf{x}}$ (and $\vec{\mathbf{n}}$) are not decorrelated, but the predicted discriminability is still valid within a scale factor, as can be verified through simulation. After some matrix algebra, we obtain

$$d^2 \propto Tr(RR^T) - Tr(R^T RM^T(MM^T)^{-1}M) \tag{15}$$

This provides a closed-form expression for the overall error as a function of the measurement matrices M and R. Finally, we wished to express this result in terms of the scaling parameters for the synthesis model and the observer. This is easily obtained from equation (15) if we assume that M and R compute local means in blocks of fixed sizes m and r, respectively, that m is an integer multiple of r, and that both m and r divide evenly into n, the length of $\vec{\mathbf{x}}$. For matrices with this structure, we can express d^2 as a function of m

$$d^2(m) \propto \begin{cases} \frac{n}{r^2}\left(1 - \frac{r}{m}\right) & m > r \\ 0 & m \le r \end{cases} \tag{16}$$

This expression has a natural continuous generalization to handle smoothly overlapping averages and non-integer ratios. The radial extent of our model pooling regions is proportional to the scaling s, so the average region size will be proportional to s^2, with a proportionality constant that depends on the shape of the region. Replacing m with s^2, and r with s_0^2, and absorbing the factor of n/r^2 into a single scale constant, gives the closed-form approximation

$$d^2(s) \approx \begin{cases} \alpha_0\left(1 - \frac{s_0^2}{s^2}\right) & s > s_0 \\ 0 & s \le s_0 \end{cases} \tag{17}$$

We empirically verified that this approximation holds for the smooth weighting functions used in our model implementation. The proportionality factor, α_0, is likely to differ for each measurement in the model. If we assume that the observer performs a weighted sum of the squared errors over the full set of measurements, then the overall error will be of the same form as that of equation (17). Notice that α_0 scales the magnitude of the squared difference, without affecting the point at which the curve first exceeds 0 (that is, $s = s_0$). Thus, when fitting the data, the gain parameter captures variability in overall performance across observers and presentation conditions. Finally, we used signal detection theory[37] to compute the probability of a correct response $P_C(s)$ in the ABX task as a function of the underlying difference $d^2(s)$

$$P_c(s) = \Phi\left(\frac{d^2(s)}{\sqrt{2}}\right)\Phi\left(\frac{d^2(s)}{2}\right) + \Phi\left(\frac{-d^2(s)}{\sqrt{2}}\right)\Phi\left(\frac{-d^2(s)}{2}\right) \tag{18}$$

where Φ is the cumulative of the normal distribution. We used the MATLAB fminsearch routine to find the values of the gain factor (α_0) and the critical scaling (s_0) that maximize the likelihood of the data (proportion correct responses for each scaling) under this model, for each subject and condition. We used bootstrapping to obtain 95% confidence intervals for the parameter estimates. We resampled the individual trials with replacement and refit the resampled data to re-estimate the parameters.

Crowding. Five observers participated in the crowding experiments (one of whom also participated in the metamer experiments). Stimuli were presented on a 13-inch flat-screen LCD monitor at a distance of 38 cm. Each observer performed two tasks, a peripheral recognition task on triplets of letters and a foveal recognition task on synthesized stimuli, similar to a previous study[24]. In the first task, each trial began with a 200-ms presentation of three letters in the periphery, arranged along the horizontal meridian. Letters were uppercase, in the Courier font, and 1 degree in height. The 'target' letter was centered at 6-degree eccentricity and the two 'flanker' letters were presented left and right of the target. All three letters were drawn randomly from the alphabet without replacement. We varied the center-to-center spacing between the letters, from 1.1 to 2.8 degrees (all large enough to avoid letter overlap). Observers had 2 s to identify the target letter with a key press (1 out of 26 possibilities, chance = 4%). Observers performed 48 trials for each spacing. For each observer, performance as a function of spacing was fit with a Weibull function by maximizing likelihood. Spacings of 1.1, 1.5 and 2 degrees corresponded to approximately 50%, 65% and 80% performance, respectively; these spacings were used to generate synthetic stimuli for the foveal task (see below). To extend our range of performance, two observers were run in an additional condition (8-degree eccentricity, 0.8 letter size, 1-degree spacing) yielding approximately 20% performance. For these observers, the same condition was included in the foveal task.

We used our mid-ventral model to synthesize stimuli matched to the letter triplets. To reduce the number of images that had to be synthesized (computational cost is high for the small scaling parameters), we synthesized stimuli containing triplets along eight radial arms, but eccentricity, letter size, font and letter-to-letter spacing were otherwise identical. For each image of triplets, we generated nine different synthetic stimuli: three different spacings (1.1, 1.5 and 2 degrees) for each of three different model scalings (0.4, 0.5 and 0.6) centered roughly around the average critical scaling estimated in our initial metamer experiment. We synthesized stimuli for 56 unique letter triplets; letter identity was balanced across experimental manipulations. On each trial of the foveal recognition task, one of the triplets from the synthesized stimuli was presented for 200 ms and the observer had 2 s to identify the middle letter. The observer saw each unique combination of triplet identity, spacing and scaling only once. Trials with different spacings were interleaved, but the three different model scalings were performed in separate blocks (with random order).

49. Schall, J.D., Perry, V.H. & Leventhal, A.G. Retinal ganglion cell dendritic fields in old-world monkeys are oriented radially. *Brain Res.* **368**, 18–23 (1986).
50. Rodionova, E.I., Revishchin, A.V. & Pigarev, I.N. Distant cortical locations of the upper and lower quadrants of the visual field represented by neurons with elongated and radially oriented receptive fields. *Exp. Brain Res.* **158**, 373–377 (2004).

© 2011 Nature America, Inc. All rights reserved.

doi:10.1038/nn.2889