Cerebral Cortex Advance Access published September 20, 2016

OXFORD

Cerebral Cortex, 2016; 1–14

doi: 10.1093/cercor/bhw282
Original Article

ORIGINAL ARTICLE

Gradual Development of Visual Texture-Selective Properties Between Macaque Areas V2 and V4

Gouki Okazawa[1,2,3,5], Satohiro Tajima[4], and Hidehiko Komatsu[1,2]

[1]Division of Sensory and Cognitive Information, National Institute for Physiological Sciences, Aichi 444-8585, Japan, [2]Department of Physiological Sciences, The Graduate University for Advanced Studies (SOKENDAI), Aichi 444-8585, Japan, [3]Center for Neural Science, New York University, New York, NY 10003, USA, [4]Department of Basic Neuroscience, University of Geneva, Geneva 1211, Switzerland, and [5]Current address: Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003, USA.

Address correspondence to Gouki Okazawa. Email: okazawa@nyu.edu

Abstract

Complex shape and texture representations are known to be constructed from V1 along the ventral visual pathway through areas V2 and V4, but the underlying mechanism remains elusive. Recent study suggests that, for processing of textures, a collection of higher-order image statistics computed by combining V1-like filter responses serves as possible representations of textures both in V2 and V4. Here, to gain a clue for how these image statistics are processed in the extrastriate visual areas, we compared neuronal responses to textures in V2 and V4 of macaque monkeys. For individual neurons, we adaptively explored their preferred textures from among thousands of naturalistic textures and fitted the obtained responses using a combination of V1-like filter responses and higher-order statistics. We found that, while the selectivity for image statistics was largely comparable between V2 and V4, V4 showed slightly stronger sensitivity to the higher-order statistics than V2. Consistent with that finding, V4 responses were reduced to a greater extent than V2 responses when the monkeys were shown spectrally matched noise images that lacked higher-order statistics. We therefore suggest that there is a gradual development in representation of higher-order features along the ventral visual hierarchy.

Key words: image statistics, macaque, material perception, texture perception, ventral visual pathway

Introduction

Visual processing of texture is pivotal for perception of the materials and surface properties of objects. We effortlessly recognize the material (e.g. wood, metal) or the condition of an object (e.g. rusted tools, rotten food), which are associated with specific textures on the object's surfaces (Adelson 2001). Indeed, many recent psychophysical studies have suggested that humans make use of various textural statistics of images to perceive materials or surface properties (for review, see Fleming 2014). Since complex visual scenes become visually indistinguishable when their textural statistics are equated (Freeman and Simoncelli 2011), understanding the neural processing of texture should provide valuable information on the mechanisms of visual perception. Previous studies revealed that the ventral pathway gradually integrates information from V1-level representation into material or surface property representation in higher-visual areas (Hiramatsu et al. 2011; Goda et al. 2014), but it remained unclear how this transformation takes place and what representations bridge them in the midlevel areas along the pathway.

Several studies have examined representations of textural features in midlevel visual areas (Hanazawa and Komatsu 2001; Arcizet et al. 2008; El-Shamayleh and Movshon 2011; Freeman et al. 2013; Okazawa et al. 2015; Yu et al. 2015; Kohler et al. 2016). Among them, recent studies suggested that statistical image features combining V1-like filter responses (i.e. spatial-

frequency and orientation filter responses) are a possible representation of naturalistic texture in the midlevel areas (Freeman et al. 2013; Okazawa et al. 2015; Ziemba et al. 2016). Those "higher-order features," originally described in Portilla and Simoncelli (2000), consist of various combinations of Gabor-like filter responses with different orientations, scales, and positions (see Fig. 1C). They provide nearly complete description of the textures in the sense that naturalistic textural images (e.g. the textures of barks or fabrics) can be synthesized using only those image features (Portilla and Simoncelli 2000). Freeman et al. (2013) found that neurons in V2 respond more strongly to naturalistic textures having those higher-order features than to spectrally matched phase-scrambled images (noise images) lacking them. We subsequently showed that neural selectivity for naturalistic textures in V4 could be parametrically fit by a version of higher-order features (Okazawa et al. 2015), indicating that those features are represented in mid-level areas along the ventral pathway.

Presuming that these higher-order features account for a part of the mechanisms for texture perception, a question arises as to how the representation of those features change along the ventral pathway, in particular how the areas thought to respond to those features (i.e. V2 and V4) differentially respond to them. Because the representation first emerges in V2 (Freeman et al. 2013; Ziemba et al. 2016), one plausible scenario is that V2 computes the higher-order features and sends copies to V4. Alternatively, those features may be further elaborated in V4. Resolving this question will clarify whether the higher-order features are considered to be purely "V2-level" image representation or whether they gradually develop along the course of the ventral visual pathway.

To address that issue, the present study directly compared texture selectivity and the underlying higher-order feature representation in V2 and V4 using exactly the same methods. We previously measured the texture selectivity of V4 neurons by adaptively exploring their preferred stimuli from among thousands of naturalistic textures (Okazawa et al. 2015). We then linearly fit the neural responses using the higher-order features. Here, we applied the same stimuli and analyses to V2 for direct comparison. We found that, while both of the areas exhibit selectivity for higher-order features, V4 responds more strongly to higher-order features than V2. This suggests a gradual development of higher-order feature representation along the ventral pathway.

Materials and Methods

We recorded neurons in area V2 and V4 in 2 female macaque monkeys ("SI" and "EV," *Macaca fuscata* weighing 5.3–6.2 kg). All procedures for animal care and experimentation were in accordance with the U.S. National Institutes of Health *Guide for the Care and Use of Laboratory Animals* and were approved by the Institutional Animal Care and Use Committee of the National Institute of Natural Sciences, Japan. Details of the methods, including visual stimulus generation and analyses, are also described elsewhere (Okazawa et al. 2015). All procedures used were identical for V2 and V4, as described below.

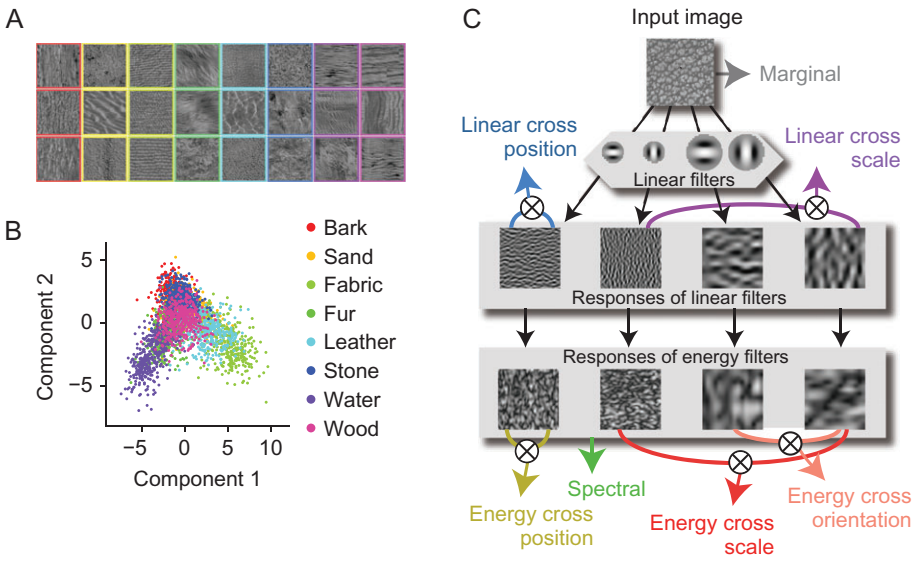


Figure 1. Stimuli used in the experiments and parameters defining the stimuli. The figures are adopted from Okazawa et al. (2015). (A) Examples of stimuli used in the experiment. We created 10 355 images and explored neurons' preferred textures by iteratively sampling a fraction of the stimuli in set. The stimuli originated from 8 material categories. The panel shows 3 example stimuli from each category. Colors of the edges surrounding the textures represent individual material categories corresponding to the dot colors in B. (B) Sampling space for adaptive exploration of textures. We projected all 10 355 stimuli onto 7-dimensional sampling space using Fisher's LDA. The panel depicts the first 2 dimensions. Each dot represents 1 image, and the dot colors represent material categories. (C) Model describing the image features. These features were used to synthesize images (Portilla and Simoncelli 2000). The total number of parameters in the original model was 740, which could be classified into 7 distinct groups of statistical features (marginal [13], linear cross position [125], linear cross scale [96], energy cross position [400], spectral [18], energy cross scale [48], energy cross orientation [40]; numbers in the brackets indicate the numbers of parameters in each group). The uppermost image represents the input texture. "Marginal" include pixel-level image statistics such as mean luminance. The image is convolved with Gabor-like filters to generate "responses of linear filter." From those, the model computes correlations between spatially neighboring filter outputs ("linear cross position") and correlations between filters with neighboring scales ("linear cross scale"). "Responses of energy filters" correspond to the amplitudes of the linear filters. From those, the model extracts average amplitudes of filter outputs ("spectral"), correlations between neighboring orientations ("energy cross orientation"), between spatial neighbors ("energy cross position"), and between neighboring scales ("energy cross scale"). Note that "spectral" can be considered as V1-level representation, while others correspond to higher-order features. For the fitting of neural responses, we used a compact version of the model, which contained only 29 parameters (see "Fitting to neuronal responses using the synthesis parameters" in Materials and Methods).

Stimulus Presentation and Electrophysiology

Stimuli were presented on a cathode ray tube monitor (frame rate: 100 Hz, Totoku Electric) situated at a distance of 57 cm from the monkeys. The stimuli were presented on the monitor using a graphics board (VSG, Cambridge Research Systems) and calibrated with a colorimeter (CS200, KONICA MINOLTA). Image resolution was 800 × 600 pixels (20 pixels/degree). The monkeys fixated on a small white spot (visual angle: <0.1°) at the center of the display. Eye position was monitored using an infrared eye camera system (ISCAN), and a trial was aborted when the eye position departed from a 1.5–2.2° diameter of fixation window.

Under aseptic conditions and general anesthesia, we surgically attached 2 chambers to the skull for V2 (left hemisphere of both monkeys) and V4 (right hemisphere of both monkeys), respectively. The stereotaxic co-ordinates of the center of the V2 chamber were 11–22 mm posterior and 9–10 mm lateral. Within the chamber, we inserted a single tungsten microelectrode (200 μm in diameter, 1–2.5 MΩ at 1 kHz; FHC, Bowdoin, ME or Unique Medical, Japan) along the sagittal plane through a guide tube attached to a plastic grid. The electrode penetrated the dura matter, went through V1 and white matter, and approached V2 in the lunate sulcus from its deeper layer. We confirmed the recording site using magnetic resonance imaging images, receptive field properties (sizes and positions), and traces of spike-activities encountered during the penetration. The stereotaxic co-ordinates of center of V4 chamber were 0–5 mm posterior and 24–29 mm lateral. An electrode, inserted through a guide tube along the coronal plane, penetrated through the dura matter and directly approached V4 in the prelunate gyrus. We recorded from 78 neurons in V2 (monkey SI 23; EV 55) and 109 neurons in V4 (SI 64; EV 45). We previously used the same 109 V4 neurons to report their texture selectivity (Okazawa et al. 2015). In both monkeys, all neurons in V2 were recorded after the recording from V4. The center and size of a neuron's receptive field were determined using a small geometric stimulus (circle, square, star, triangle, or bar) that evoked the neuron's maximal response. We changed the position of the geometric stimulus in 1° intervals and determined the border of the receptive field as the position at which the visual response ceased. For 2 cells in V4 that did not respond to any of the small stimuli, only the center of the receptive field was determined using a texture (6.4° × 6.4°). During the recordings, a trial started with presentation of a fixation spot, after which textural stimuli were presented 5–6 times within each trial. Each stimulus presentation lasted 200 ms with 200-ms blank intervals. Each stimulus was repeated at least 5 times and usually 8 times.

Visual Stimulus Generation

Stimuli were generated using a texture synthesis procedure described in Portilla and Simoncelli (2000) with a program provided by the authors (http://www.cns.nyu.edu/~lcv/texture/). The program extracts a collection of higher-order statistics and spectral statistics from an image (Fig. 1C) and synthesizes a new image having nearly identical statistical parameters. During the synthesis, the algorithm starts from a white-noise image and iteratively modifies the image to match its statistics with the desired one. We performed 50 iterations and confirmed the convergence of parameters. For the synthesis, we used 4 scales and 4 orientations of Gabor-like filters to extract the statistics. For the parameters describing correlations between spatially neighboring filter outputs, correlations within a 7-pixel square were taken into account.

We generated 10 355 synthetic textures. Some of the images were synthesized from the synthesis parameters of 4170 photographs sampled from 8 material categories (bark, sand, fabric, fur, leather, stone, water, and wood; an average of 521 images for each), which were collected from commercial databases (SOZAIJITEN, Datacraft, Japan) and the Internet. The other 6185 images were generated using synthesis parameters interpolated from the above images. The interpolation was performed in a sampling space explained in the next paragraph. The purpose of interpolation was to mitigate inhomogeneous distributions of stimuli within the sampling space. For this, we computed the distances of 20 adjacent textures in each stimulus within the sampling space and interpolated stimuli if the distances were more than 3 standard deviations (SDs) of the whole distance distribution.

We projected all images onto a sampling space (Fig. 1B), which enabled the adaptive sampling of the stimuli within the space. To generate a sampling space, we normalized individual synthesis parameters (740 parameters) across all natural texture images, denoised them using principal component analysis, which reduced the dimensions to 300, and finally projected them into a 7-dimensional space using Fisher's linear discriminant analysis (LDA). LDA finds the linear subspace that maximally separates different categories (Bishop 2006). As mentioned above, we added 6185 points by interpolating the parameters of 4170 textural images within the sampling space. No category label was assigned to those interpolated images. The size of the images was 128 × 128 pixel (corresponding to 6.4°). They were all gray scale. The mean and SD of the luminance histogram were, respectively, equalized to 15 and 6 cd/m^2 to avoid the effects of these low-level factors. We presented the images on a gray background (10 cd/m^2).

Adaptive Sampling Procedure

To efficiently find the neurons' preferred textural stimuli, we applied an adaptive sampling procedure (Yamane et al. 2008). We first randomly selected 50 textures (the first generation) from the 10 355-image set and recorded the single-cell responses they elicited. Then, in subsequent generations, we selected stimuli among neighbors of ancestor stimuli selected from earlier generations based on the ranks of the elicited firing rates: 13 from the top 10% of stimuli, 10 from the next 10–24%, 5 from the 24–44%, 5 from the 44–70%, and 5 from 70–100%. Each subsequent generation also included 12 new, randomly selected stimuli. We included the random sampling to keep the process from falling into local minimums. We repeated this procedure at least 5 times and at most 10 times to record neuronal responses to 250–500 textures.

Using this approach, we were able to present multiple stimuli that evoked strong firings and thus adequately characterize the neuron's tuning properties. The rationale behind this adaptive sampling procedure is as follows. If there is only a small, predetermined stimulus set, it is less likely to observe strong responses, and one will end up recording only from neurons responsive to the predetermined stimuli, which would result in a biased sampling of the neural population. Furthermore, if a neuron responded to only a fraction of the images, it would be challenging to fit a model to the data, since the information available from the small number of effective stimuli is limited. Obtaining strong neural responses, however, is not trivial when the potential dimensionality of the stimulus space is large and

there is little prior information about the neurons' tuning. The adaptive sampling procedure should mitigate these issues by dynamically adjusting the stimulus set based on obtained neural responses. And in fact, we succeeded in obtaining higher-firing rates using this method as is described in the Results section (see "Texture selectivity in V2 and V4") and in our earlier report (Okazawa et al. 2015).

Analysis of Neuronal Firing Rates

Each neuron's mean firing rates were defined as the firing rates averaged across a 200-ms period beginning 50 ms after stimulus onset and ending 50 ms after offset for both V2 and V4. We subtracted the baseline activity, defined as the firing rates during the 300 ms before the first stimulus onset averaged across all trials in each generation of adaptive sampling.

We characterized texture selective properties using several measures. As a measure of a neuron's ability to discriminate textural stimuli, we used discrimination index (Prince et al. 2002; Sanada et al. 2012):

$$\text{discrimination index (DI)} = \frac{R_{max} - R_{min}}{R_{max} + R_{min} + 2\sqrt{SSE/(M-N)}}, \quad (1)$$

where R_{max} and R_{min} are the maximum and minimum average firing rates across all stimuli. We used square roots of firing rates instead of the raw firing rates to roughly equalize the variances of firing rates (Prince et al. 2002). SSE is the sum of squared error of the responses to individual stimulus presentations around the average firing rates, M is the total number of stimulus presentations, and N is the number of stimuli. Thus, the discrimination index indicates the difference between the maximum and minimum firing rates normalized by their sum and variability of firing rates. The index is helpful for comparing different brain regions, as the variability of responses may differ across areas. As a simpler control, we also used a simpler version of the discrimination index that did not incorporate the variability of the firing rates (i.e. difference between the maximum and minimum firing rates divided by their sum). The discrimination indices range from zero to one, and larger indices indicate better discriminability.

We also evaluated sharpness of neuronal tuning using sparseness index (Vinje and Gallant 2000):

$$\text{sparseness index (SI)} = \left[1 - \left(\sum_i \frac{R_i}{N}\right)^2 / \sum_i \frac{R_i^2}{N}\right] / \left(1 - \frac{1}{N}\right), \quad (2)$$

where R_i is the firing rate elicited by stimulus i and N is the number of stimuli. The value ranges from zero to one, and larger sparseness indices indicate sharper selectivity. Because it has been suggested that this index is vulnerable to changes in other firing properties such as minimum firing rates and Poisson properties of spike production (Lehky et al. 2005; Rust and DiCarlo 2012), we also computed the entropy (S_E) as another measure of the sharpness of tuning that provides information about the shape of the response distribution invariant to the mean response. Entropy was defined as

$$S_E = 2.074 + \sum_{j=1}^{M} p(R_j)\log_2(p(R_j))\Delta R, \quad (3)$$

where responses to N images were placed in M bins, where M is calculated as the square root of the number of images N (Lehky et al. 2005). M bins divided the entire range of the responses with an equal interval ΔR. The value 2.074 corresponds to the entropy of a Gaussian distribution with unit variance, which is the maximum value of possible entropies when the variance is fixed. Before computing entropy, we rescaled the firing rates to have unit variance. S_E ranges from 0 to 2.074, with larger values indicating sharper selectivity. We also computed a "corrected sparseness index," which amends the bias originating in the Poisson property of spike production in a cell (Rust and DiCarlo 2012). Because the Poisson distribution is positively skewed, responses to some stimuli can accidentally exhibit a large firing rate, which results in an overestimation of the sparseness index. As described in Rust and DiCarlo (2012), we assumed that a neuron's rank-order response curve approximately follows an exponential form: $R(x) = Ae^{-\alpha x}$, where x is a stimulus rank sorted by the evoked firing rates (x can take the value from 1 to the number of stimuli). We then computed the sparseness index of the estimated firing rates $R(x)$ instead of the actual firing rates. To obtain the free parameters A and α, we used a maximum likelihood estimation, presuming that the recorded firing rates for stimulus x were drawn from Poisson distributions of $R(x)$.

Fitting to Neuronal Responses Using the Synthesis Parameters

The purpose of this fitting analysis is to examine how well neural activities could be explained using the spectral and higher-order statistical parameters used in the synthesis algorithm (Fig. 1C). The total number of parameters in the original model is 740, which could be classified into 7 distinct groups of statistical features (marginal [13], linear cross position [125], linear cross scale [96], energy cross position [400], spectral [18], energy cross scale [48], energy cross orientation [40]; numbers in the brackets indicate the numbers of parameters in each group). From an input texture, "marginal" statistics are computed, which include mean, SD, skewness, and kurtosis of luminance histogram. The image is convolved with Gabor-like filters to generate "responses of linear filter" (Fig. 1C). From those, the model computes correlations between spatially neighboring filter outputs ("linear cross position") and correlations between filters with neighboring scales ("linear cross scale"). Then, "responses of energy filters" are obtained by computing the amplitudes of the linear filters. From those, the model extracts average amplitudes of filter outputs ("spectral"), correlations between neighboring orientations ("energy cross orientation"), between spatial neighbors ("energy cross position"), and between neighboring scales ("energy cross scale"). Among these 7 groups of statistics, "spectral" can be considered as "V1-level representation" because they correspond to the average responses of energy filters of different orientations and scales, while other groups are considered as "higher-order features." Among the "marginal" statistics, mean and SD of luminance are matched in our stimulus set and we only used skewness of luminance histogram. This parameter was categorized into higher-order features.

Because the number of parameters in the synthesis model is large (740), which is unfavorable for the fitting analysis, we generated a compact version with 29 parameters that preserves the core features of the model. The sampling space we generated for the adaptive procedure was also derived from the synthesis model (Fig. 1B), but we did not directly use them for fitting because we found that the space did not have a sufficient power to predict neural responses compared with the 29 parameters.

We used 2 strategies to generate a compact version of the synthesis parameters. First, because parameters extracted from neighboring filters are correlated, we averaged across the values of neighboring filters (originally 4 scales × 4 orientations) and compressed them into 2 scales × 2 orientations. Second, we applied principal component analysis for highly redundant parameter groups such as "linear cross position" and "energy cross position" statistics (Fig. 1C). Although they have more than hundreds of parameters, the first 3 to 4 principal components explain more than 80% of the variances. We therefore retained only those principal components. We performed principal component analysis only within these parameter groups but not across the whole parameter set so as not to mix different parameter across groups in the model.

We used L1-penalized linear least-squares regression (known as lasso) (Tibshirani 1996) to fit the firing rates of neurons elicited by 250–500 textures. The regression minimizes the following loss function L:

$$L = \frac{1}{N}\sum_{i=1}^{N}(nFR_i - PS_i \cdot W)^2 + \lambda |W|, \quad (4)$$

where N is the number of stimuli, nFR_i and PS_i are the observed firing rate normalized within the cell and synthesis parameters for image i, W is the fitting weights, and λ is the regularization coefficient. We cross-validated the performance of the fit by partitioning the data into training (randomly selected 90% of presented textures) and test (remaining 10%) sets. Correlation coefficients (Fig. 5A) were computed between the actual neural responses in the test set and neural responses predicted from the fitting result. We repeated this procedure 10 times with different combinations of training and test sets and then averaged across their obtained correlation coefficients (so-called 10-fold cross-validation). We also chose λ in eq. 4 using cross-validation (5-fold) within the training set so as to provide the optimal fitting performance within the training set. The statistical significance of the fitting was tested using a permutation test in which we shuffled the combinations among the textures and firing rates to obtain the distribution of correlation coefficients when there is no relationship between stimuli and firing rates. For the final estimation of fitting weights (displayed in Fig. 5B), we used the whole data set without the cross-validation. For the further analyses using the weights, we normalized the weights within the cells so that their mean becomes one.

To statistically test the difference in fitting weights between V2 and V4, we performed a permutation test. We first computed the population average of the fitting weights for V2 and V4 (Fig. 5C) and quantified their difference using the sum of squared error: $SSE = \sum_i (V2_i - V4_i)^2$, where $V2_i$ and $V4_i$ indicate the average weights of statistical group i. During the permutation test, we computed the SSE after randomly shuffling neurons between V2 and V4 while preserving the number of neurons in each area. We repeated this procedure 100 000 times and computed the probability that the SSE of the shuffled data exceeded the actual SSE.

To gain insight into how stimulus selectivity emerges over time, we also analyzed the temporal dynamics of the firing rates and selectivity for image statistics. To draw peristimulus time histograms (PSTHs; Fig. 7A), we counted the numbers of spikes in nonoverlapping 10-ms windows and averaged across all stimuli and neurons. The time course of selectivity for image statistics (Fig. 7C) was obtained by performing the above fitting analysis for firing rates in each time bin.

Results

Texture Selectivity in V2 and V4

We analyzed 78 neurons in V2 (monkey SI 23; EV 55) and 109 neurons in V4 (SI 64; EV 45). The eccentricities of their receptive field centers were comparable (V2: 5.1 ± 1.3°; V4: 6.0 ± 2.4°). We tested the responses of every neuron encountered during the experiment and recorded all neurons that responded to any of our textural stimuli (144 neurons in V2; 225 neurons in V4). Of these, 138 neurons in V2 (96%) and 214 neurons in V4 (95%) showed significant texture selectivity ($P < 0.05$; Kruskal–Wallis test). We adaptively explored their preferred stimuli among 10 355 synthetic textures generated prior to the experiment. In this adaptive sampling procedure, we first recorded the responses of a single cell to 50 randomly selected textures. We then chose next textures such that spaces around the preferred textures were densely sampled in the predefined texture space (Fig. 1B). Among the recorded texture selective cells, we analyzed neurons that could be recorded at least 5 generations of the adaptive sampling (i.e. more than 250 stimuli; 78 neurons in V2; 109 neurons in V4). On average, we sampled responses to 297 ± 79 stimuli for each V2 cell and 357 ± 83 for each V4 cell.

All analyzed neurons in both V2 and V4 showed significant texture selectivity ($P < 0.001$, Kruskal–Wallis test). These neurons vigorously responded to some stimuli but not to others (Fig. 2). One V2 neuron, for example, responded strongly to wavy textures, such as the surface of water (Fig. 2A; "V2 cell 1"), while another responded to mesh textures such as leather (Fig. 2A; "V2 cell 2"). Similarly, individual V4 neurons preferred various types of textures such as those of wood (Fig. 2A; "V4 cell 1") and fabric (Fig. 2A; "V4 cell 2"). On the whole, visual inspection of their preferred textures revealed no clear differences between V2 and V4 (more examples in Fig. 2B). Quantitatively, there was no significant difference in the strength of the texture selectivity between the areas when tested using a discrimination index (V2, 0.72 ± 0.12; V4, 0.74 ± 0.08; $P = 0.50$, Mann–Whitney U test; see eq. 1 in Materials and Methods), although we found several differences in response properties (e.g. mean evoked firing rates, sparseness of responses) between the areas, which we elaborate on later ("response characteristics" section in Results).

Our adaptive sampling procedure helped us to efficiently collect neural responses to textures that evoked high firing rates in each neuron. For each generation of sampling, we chose a new set of textures such that many were descendants of textures that evoked strong responses, while others were randomly selected textures (Fig. 3A). We found that at the population level in both V2 and V4, the textures chosen as descendants of those that elicited strong responses were indeed more likely to elicit stronger responses than either the descendants of textures that elicited only weak responses or randomly selected stimuli (Fig. 3B; group A vs. B–E, R: $P < 0.001$; Wilcoxon signed rank test). This indicates that the sampling procedures succeeded in collecting more responses than a random sampling.

Tuning for Image Statistics

As in our earlier study (Okazawa et al. 2015), we examined how well the neurons' selectivity for textures can be explained by sensitivity for spectral and higher-order image statistics. Because the higher-order statistics in the texture synthesis model consisted of a very large number of parameters (740), we first reduced that number to 29 using dimensionality reduction

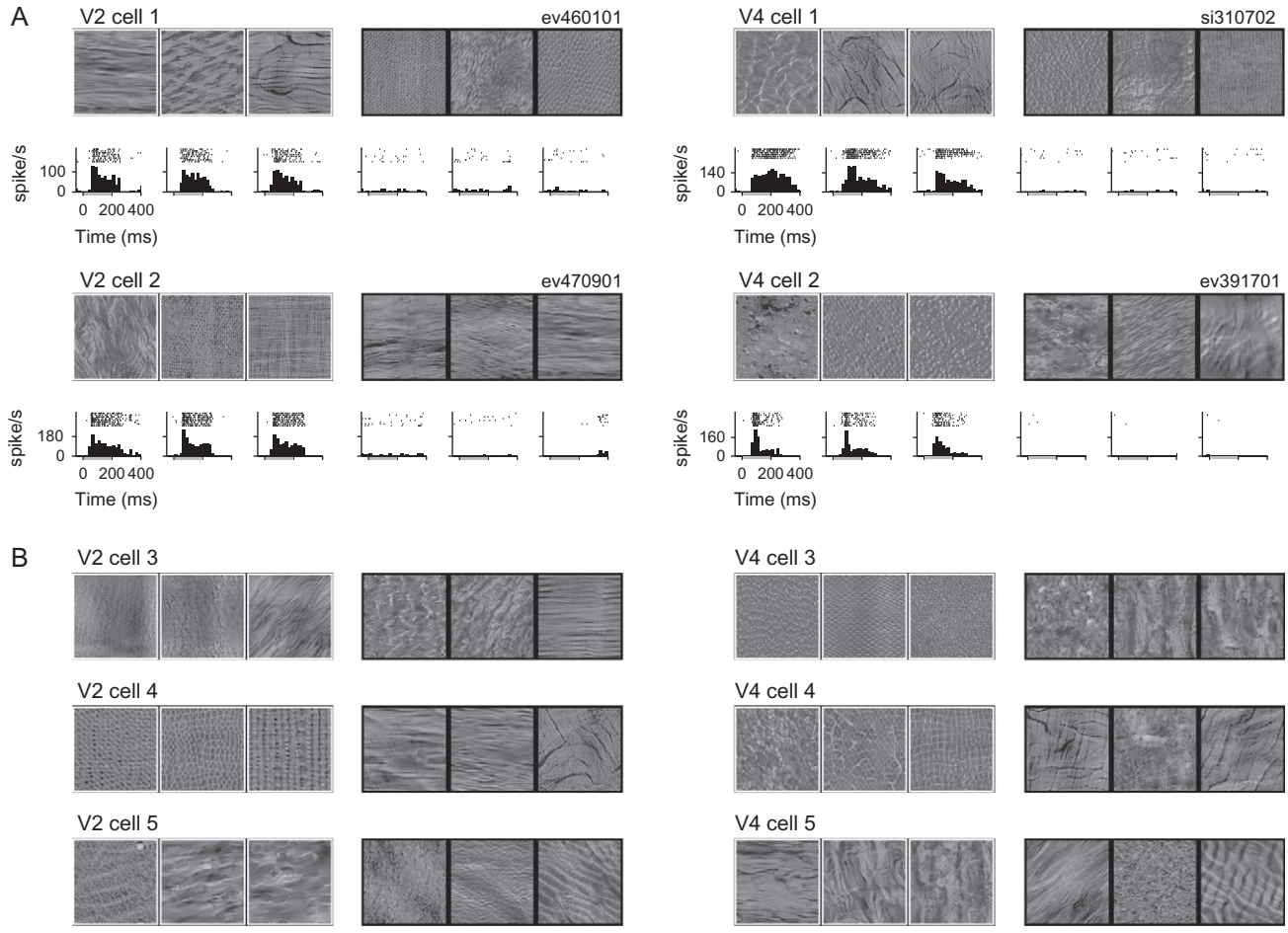


Figure 2. Texture selectivity of neurons in V2 and V4. (*A*) Four example cells from V2 and V4. The panels depict the 3 most preferred (surrounded with white edges) and least preferred (surrounded with black edges) textures with the corresponding PSTHs and raster plots of cells' responses. A gray horizontal bar underneath each PSTH indicates the stimulus presentation period. (*B*) Texture selectivity of 3 more example cells each from V2 and V4. The shown textures are the 3 most (white edges) and the least (black edges) preferred stimuli for each cell.

techniques (see Materials and Methods). These 29 parameters included the average powers of spatial-frequency and orientation filters (spectral statistics; green label in Fig. 1C) considered as V1-level representation, as well as combinations of those filter responses with different scales, orientations, and positions (higher-order features; other colors in Fig. 1C).

We found that texture-selective responses were associated with the presence of specific image statistics in both V2 and V4. To gain information about the relationships between textures and the statistical parameters, the most preferred 2 textures and their statistical parameters for example neurons (same as those shown in Fig. 2A) are depicted in Fig. 4. V2 cell 1 (Fig. 4, upper left) preferred wavy textures that had strong horizontal low-frequency spectral components, as depicted in the "spectral" graph, which indicates the amplitudes of each spatial-frequency and orientation component. The images did not have strong higher-order features such as "energy cross scale." By contrast, the preferred textures of V4 cell 1 (Fig. 4, upper right) had strong "energy cross scale" correlations (the correlation of energy filters between 2 scales) between high- and low-frequency vertical filters due to the presence of sharp edges in the images. The 2 images did not have consistent spectral features. V2 cell 2 (Fig. 4, lower left) preferred textures that contained "energy cross orientation" correlations (the correlation of energy filters between 2 orientations) between vertical and horizontal filters, which are frequently observed in images containing mixtures of vertical and horizontal lines. Similarly, V4 cell 2 (Fig. 4, lower right) preferred textures containing "energy cross orientation."

To quantify the relationship between neurons' responses and the statistical parameters of textural images, we fit the firing rates to all presented textures using a regularized (L1-penalized) linear regression (Tibshirani 1996) of the 29 parameters, including spectral and higher-order parameters. The regularized linear regression is less likely to fall into overfitting of data than typical linear regression, especially when the number of fitting parameters is large. To comply with the criterion used in our earlier study (Okazawa et al. 2015), we excluded a fraction of cells that responded sparsely to the textural stimuli (sparseness index > 0.75; see Materials and Methods), since such a sparse response is unfavorable for fitting analyses (remaining cells: V2, 64 cells (82%); V4, 90 cells (83%)). We obtained similar results when those cells were not excluded, as is described in Supplementary Figure 1. To evaluate the fitting performance, we divided all stimuli presented to a cell into training and test groups; the fitting weights were then computed using the training set (90% of stimuli), and correlations between predicted and observed firing rates were examined using the test set (10% of

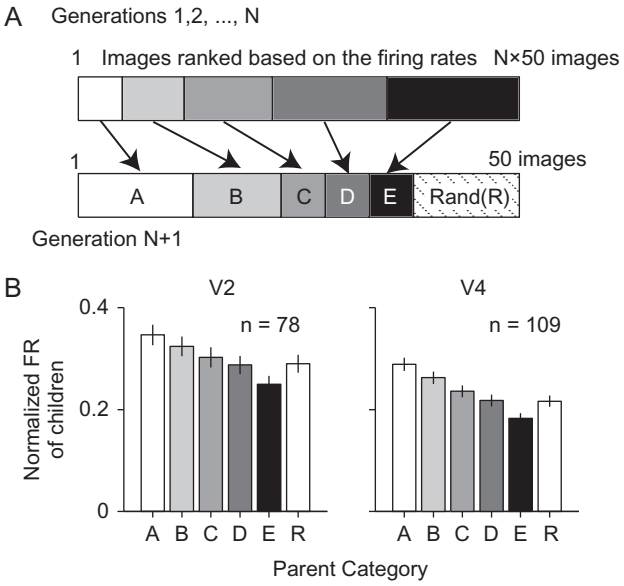


Figure 3. Efficient collection of effective stimuli using an adaptive sampling procedure. (*A*) The schema of adaptive sampling. To generate a new set of stimuli ("Generation *N*+1"), we sorted the firing rates elicited by all stimuli presented so far ("Generation 1, 2, … , *N*") and chose textures in the sampling space so that more stimuli were selected from neighbors of effective textures. We also included 12 randomly sampled textures ("Rand (R)") in each generation. The figure is adopted from Okazawa et al. (2015). (*B*) Firing rates elicited by textures sorted based on the textures' parent categories (A–E and R corresponds to those in panel *A*). Firing rates were normalized in each cell. We averaged neural responses obtained in all generations except the first one, which does not have parent categories. Firing rates tended to be large when the parent stimuli elicited strong responses. Error bars indicate the standard error of the mean (SEM) across cells. The V4 data were adopted from Okazawa et al. (2015).

stimuli). This means the fitting performances were examined using data independent of those used to estimate the weights. For many neurons in V2 and V4, we obtained statistically significant correlations between the actual firing rates and the firing rates predicted by the model (Fig. 5*A*). In V2, 56 of 64 neurons (88%) showed correlations significantly above the chance level (r averaged across 64 cells = 0.45; $P < 0.05$, permutation test). In V4, 83 of 90 neurons (92%) showed a significant fit (r averaged across 90 cells = 0.46). These fitting performances were statistically indistinguishable ($P = 0.89$, t-test using Fisher's Z-transformation). The fitting explained 29 ± 17% (V4) and 31 ± 14% (V2) of the explainable variance in the firing rates computed by subtracting the trial-by-trial variance from the whole variance (Pasupathy and Connor 2001). We also examined the fit of neural responses using texture models other than the synthesis parameters, but we found no models that outperformed our texture model (Supplementary Figure 2) in terms of the fitting performance obtained using the same fitting procedure.

The linear fitting weights illustrated in Fig. 5*B* indicate that individual cells have weights on various different parameters. Although the significant fitting performances do not mean that our texture model is the "true" underlying model of V2 and V4, it is worth inspecting how the model weighted parameters to fit the neural responses. Any differences in weighted parameters between V2 and V4 will give us insight into how the 2 areas differentially respond to textures. In the color map shown in Fig. 5*B*, the brightness of the dots indicates the amplitude of weights (normalized within individual cells) and the colors indicate the groups of statistical features, which correspond to the label colors in Fig. 1*C*. Cells were sorted based on the most preferred group of statistical features. Individual columns contain a few bright dots, meaning that each cell has weight on a few parameters. For example, some cells exclusively had weight on spectral features (columns that contain bright green dots), which are considered to be V1-level representation, while others had weight on other higher-order features (columns that contain bright colors other than green). These weight maps were comparable between V2 and V4, though V2 appears to contain slightly more neurons having weights on spectral features (green dots), while V4 appears to contain a larger number of neurons having maximum weight on higher-order energy features (yellow, orange, and red dots at the right bottom corner of the diagram).

At the population level, we found similar but slightly different weights for statistical parameters in V2 and V4. To quantify the population-level selectivity, we averaged all neurons' weights for each group of statistical features (Fig. 5*C*). Both V2 and V4 showed stronger weights for spectral features as well as higher-order statistics such as those called "energy cross orientation" or "energy cross position." The population-level preferences for statistical features looked similar in V2 and V4, though V2 appears to have relatively stronger weights for the spectral statistics, and V4 appears to have stronger weights for energy statistics. A statistical test revealed significant differences in the distributions of weights between V2 and V4 ($P = 0.006$; permutation test; see Materials and Methods). Significant differences in individual parameters were seen for "spectral," "energy cross orientation," "energy cross position," and "energy cross scale" statistics ($P < 0.05$, Mann–Whitney U test without Bonferroni correction). With Bonferroni correction, "energy cross position" showed a significant difference ($P = 0.005$). That pattern—greater weights for energy statistics and smaller weights for spectral statistics in V4—was consistently observed in each monkey (Fig. 5*D*).

To quantify the relative balance of weights between the spectral and higher-order parameters, we computed a higher-order ratio, which is the mean amplitude of weights for higher-order statistics divided by the sum of the mean amplitudes of spectral and higher-order statistics. When we sorted neurons according to the amplitudes of the higher-order ratios (Fig. 5*E*), we found that the V4 distribution consistently exceeded that of V2 ($P = 0.035$; Mann–Whitney U test). Taken together, these findings indicate that, while the weights of V2 and V4 are generally comparable, there is a relative imbalance in their selectivity for V1-level spectral and higher-order features.

While the eccentricities of the receptive fields are similar between V2 and V4 (V2: 5.1 ± 1.3°; V4: 6.0 ± 2.4°), there was a large difference in the average size of the receptive fields (V2, 2.4 ± 1.2°; V4, 6.1 ± 3.1°). To examine the effect of the smaller receptive fields in V2, we recalculated the statistical parameters of textural images after cropping the central parts so that the stimulus size was the same as the receptive field size of each V2 neuron. We then refit the neural data using those recalculated parameters. This revealed that the amplitudes of the fitting weights for each group of statistics were little affected by the size of patches (Pearson's correlation coefficient with Fig. 5*C*, V2 data: 0.99). We also found that there is no significant correlation between receptive field size and the amplitude of the higher-order ratio in either area or across areas (V2: $r = -0.095$, $P = 0.49$; V4: $r = -0.11$, $P = 0.33$; both: $r = 0.008$, $P = 0.92$). The absence of correlation indicates that the differences in selectivity between the 2 areas cannot be accounted for by the receptive field sizes.

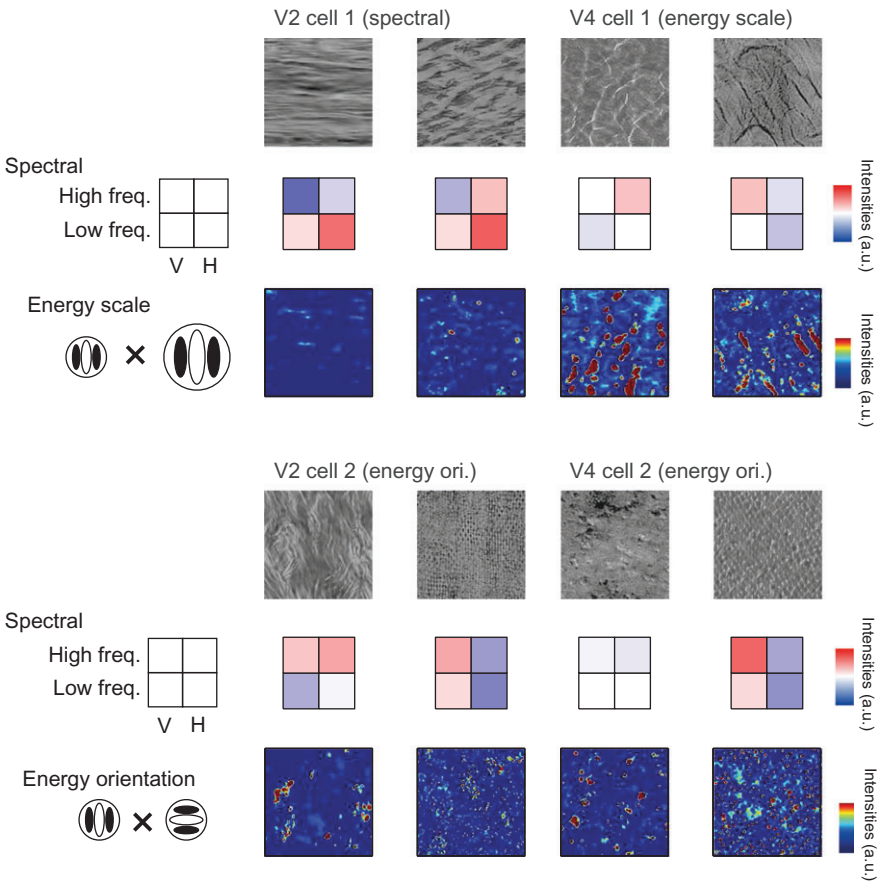


Figure 4. Relationships between textures and statistical parameters. The 2 most preferred textures of the 4 example cells area depicted with their statistical parameters. The 4 cells (2 V2 cells and 2 V4 cells) correspond to those in Fig. 2A. Each neuron's preferred statistics is indicated in the parentheses, which was obtained using the fitting analysis described in Fig. 5. For each texture, we depicted "spectral" parameters and "energy scale" or "energy orientation" parameters. The spectral parameters consist of 4 components: high-/low-frequency vertical (V)/horizontal (H) amplitudes. Colors indicate the amplitudes of those components normalized across all texture stimuli. "Energy scale" and "energy orientation" parameters indicate co-occurrence of responses between 2 filter outputs with different scales and orientations, respectively. We depicted the locations in the images that have such co-occurrences using a color scale. Actual parameters used in the fitting analysis were the average of those co-occurrences across locations.

We also fitted the neural responses using only the spectral statistics, which enabled us to examine how well the spectral parameters alone account for the neural responses. The fitting performances for neurons in both V4 and V2 significantly declined (V4: $r = 0.35$, $P < 0.001$, V2: $r = 0.40$, $P < 0.001$; Wilcoxon signed rank test; the results of 10-fold cross-validation). The differences in fitting performances are unlikely to be explained solely by the numbers of parameters as we tested the performances using an independent test set. Because the V2 fitting performance deteriorated to a lesser extent, the fitting performances using the spectral statistics were significantly better in V2 than in V4 ($P = 0.049$, Mann–Whitney U test). We also examined other models representing spectral features with larger numbers of parameters (Supplementary Figure 2; "gist," "spectral RF"). These models tended to fit neural responses better for V2 than for V4 although none outperformed our texture model. These results are consistent with the idea that V2 neurons exhibit stronger preference for spectral parameters.

Neural Responses to Spectrally Matched Noise Images

To obtain independent support for the observations in the fitting analyses, in a subset of cells, we performed a control experiment using noise images that lacked higher-order feature (35 cells in V2 and 83 cells in V4). We chose 5 textures that were evenly selected from the cell's preferred and nonpreferred stimuli and generated corresponding control ("noise") images by randomizing the phases of the spatial-frequency components for each original texture in the Fourier space (Fig. 6A shows example images). Those noise images had the same spectral statistics as the original textures, but their higher-order statistics were greatly deteriorated. Indeed, when we compared the amplitudes of the higher-order parameters between all the textures and the noise images used, significant differences were found in "marginal" and all "energy" statistics ($n = 590$ textures; $P < 0.05$, Wilcoxon signed rank test). Thus, differences between the responses to the textures and those to the corresponding noise images can be a good measure of the effect of those higher-order statistics on neuronal responses.

Consistent with the fitting analyses, we found significant differences between responses to the textures and those to the noise stimuli in both V2 and V4 (Fig. 6B; V2: $P = 0.002$, V4: $P < 0.001$; Wilcoxon signed rank test). This indicates that removing higher-order statistics reduced neural responses in both areas. To determine which area showed the larger difference between responses to textures and noise, we computed a modulation index, which is the difference in normalized firing rates between the textures and noise images divided by their

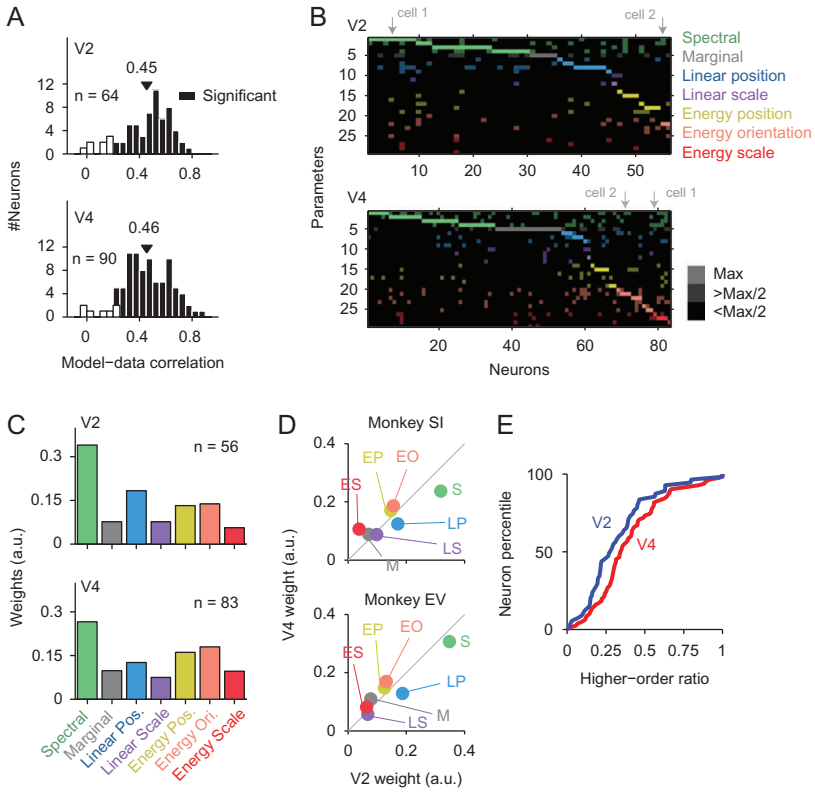


Figure 5. Result obtained by fitting neuronal responses using the spectral and higher-order image statistics depicted in Fig. 1C. (*A*) Distribution of the fitting performance. The fitting performance was evaluated using Pearson's correlation coefficients between neuronal responses and responses predicted by the model (the abscissa). Independent data were used for fitting and computation of the correlation ("10-fold cross-validation"; see Materials and Methods). The black bars indicate the neurons that showed significant fit ($P < 0.05$, Permutation test). The arrowhead denotes the average correlation coefficient across cells. The V4 data were adopted from Okazawa et al. (2015). (*B*) Linear fitting weights of all neurons depicted using colors and pixel brightness. We used 29 parameters to fit individual neuron's responses, and the vertical axis corresponds to those 29 parameters. The brightness of colors indicates the strength of the absolute amplitude of the weights (the brightest: maximum weight, the middle: weights more than half maximum, the darkest: weights less than half maximum). Colors indicate different groups of image statistics (see Fig. 1C). The neurons were sorted based on their most preferred groups of image statistics. Gray arrows point to the example cells shown in Fig. 4. (*C*) Weights averaged across cells. The weights were normalized within each cell before the averaging. Bar colors indicate groups of image statistics and correspond to colors in Fig. 1C. (*D*) Normalized weights for each group of statistical parameters averaged across cells in V2 and V4 for each monkey. The gray oblique line indicates unity. Colors represent different groups of statistical parameters: S, spectral; M, marginal; LP, linear position; LS, linear scale; EP, energy position; EO, energy orientation; ES, energy scale. (*E*) Distributions of higher-order ratios for V2 ($n = 56$) and V4 ($n = 83$). The higher-order ratios were computed by dividing the mean amplitude of the weights for higher-order parameters by the sum of the mean amplitudes for spectral and higher-order parameters. We sorted neurons based on the amplitudes of the higher-order ratios and displayed them as cumulative histograms. Line colors denote V2 (blue) and V4 (red).

sum. The modulation index was larger in V4 than in V2 (Fig. 6C; $P = 0.008$; Mann–Whitney U test), which is consistent with the idea that V4 responds to higher-order features more strongly than V2. We also examined how well the responses to noise stimuli could be explained using the fitting weights obtained above. For each cell, we computed the synthesis parameters of the presented noise images and calculated the predicted responses to those images using the neuron's fitting weights. The average correlation coefficients between the observed and predicted responses to the 5 noise images were 0.48 for V2 ($n = 35$) and 0.47 for V4 ($n = 83$), which are similar to those obtained in the fitting analysis. This result further supports the plausibility of fitting analysis.

Dynamics of Tuning

We next sought to analyze how the selectivity for spectral and higher-order image features changes over time in V2 and V4 because the temporal dynamics of stimulus selectivity provides insight into the mechanisms of sensory processing (Hegdé and Van Essen 2004; Brincat and Connor 2006; Hegdé 2008). PSTHs of neural firing rates displayed earlier onset in V2 than in V4 (Fig. 7A). At the population level, the responses in V2 began at around 30 ms and peaked at around 70 ms, while those in V4 began at around 50 ms and peaked at around 120 ms. We binned spikes using 10-ms nonoverlapping time windows and conducted the fitting analysis described above for each of these time bins. The fitting performances, quantified using Pearson's correlation coefficient, exhibited time courses similar to those of the PSTHs (Fig. 7B).

The dynamics of the weights for spectral and higher-order statistics revealed characteristic differences between V2 and V4 (Fig. 7C). For the spectral parameters, V2 showed an early and sharp rise in weight followed by a continuous decline, while V4 showed a relatively weak weight. A significant difference was observed during the early period (40–90 ms; $P < 0.05$, Mann–Whitney U test). For the higher-order features, V2 again showed an earlier onset, but, in contrast to spectral parameters, V4 showed larger weights. Consequently, V4 showed significantly larger weights around 100 ms, though a difference was also observed around 60 ms due to the earlier onset in V2.

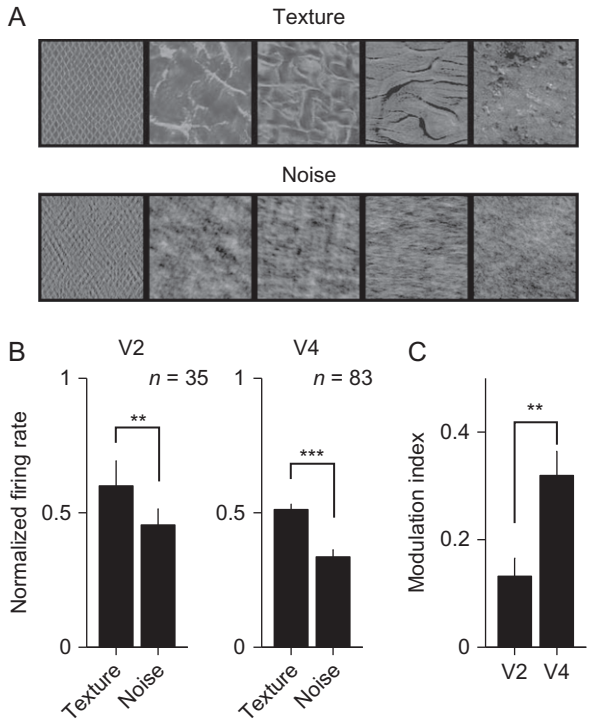


Figure 6. Control experiment using spectrally matched noise stimuli. (*A*) Example textures and corresponding noise stimuli used in the control experiment. The phase of each texture in the Fourier space was randomized to create a noise stimulus. (*B*) Normalized firing rates for textures and noises in V2 and V4. We showed 5 textures and noise images that were evenly chosen from a cell's preferred and nonpreferred textures and averaged the elicited responses. The firing rates were normalized by the response to the cell's most preferred texture. Error bars are SEM across cells. **P < 0.01, ***P < 0.001, Wilcoxon rank sum test. (*C*) We computed a modulation index for each cell, which is the difference in normalized firing rates between the textures and noise images divided by their sum. Error bars are SEM across cells. **P < 0.01, Mann–Whitney *U* test.

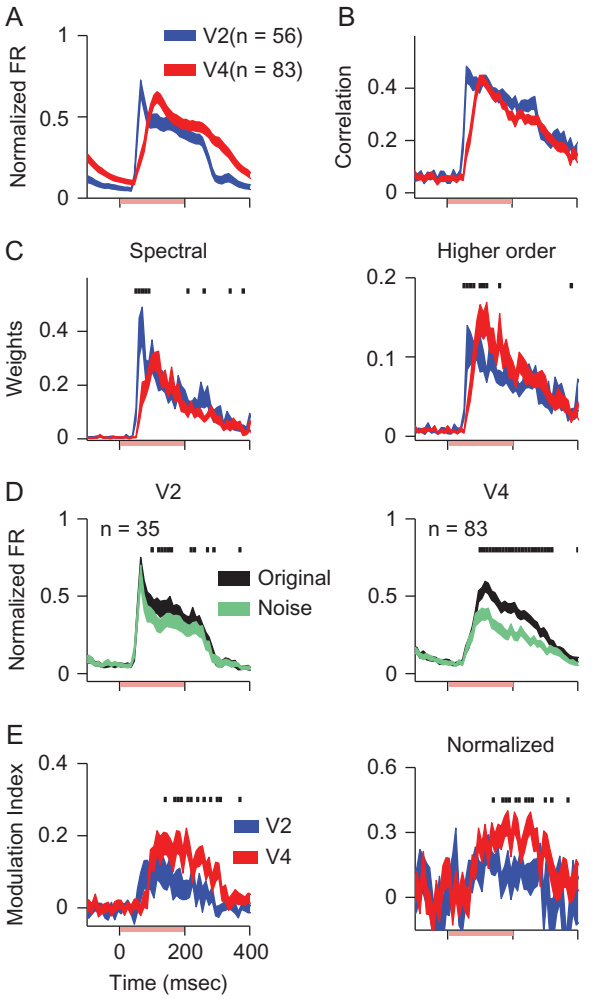


Figure 7. Analysis of the time course of responses and selectivity for image features. (*A*) PSTH averaged across the responses of all significantly fit neurons in V2 (blue; $n = 56$) and V4 (red; $n = 83$). Before averaging across cells, the firing rates were normalized such that the maximum for each cell was unity. The line thickness indicates the SEM across cells. A pink horizontal bar beneath the lines indicates the stimulus presentation period. (*B*) Time courses of the fitting performances using the 29 image statistics. The panel shows the average of all significantly fit cells. The fitting analysis was performed for firing rates within every 10-ms nonoverlapping time windows. The vertical axis corresponds to Pearson's correlation coefficients between actual neural responses and responses predicted from the fitting analysis. (*C*) Mean of weights across cells for spectral statistics (left) and higher-order image statistics (right). The black lines above the traces indicate the periods where the difference between V2 and V4 was significant ($P < 0.05$; Mann–Whitney *U* test). (*D*) PSTHs averaged across textural images (Original) and corresponding spectrally matched noise images (Noise), which have the same spectral parameters but lack higher-order statistics: left, V2; right, V4. The responses were average across the neurons tested using the noise images (V2, $n = 38$; V4, $n = 90$). Responses were normalized to the peak amplitude in each cell. The line thickness indicates SEM across cells. (*E*) Modulation index, defined as the differences in responses between textural stimuli and corresponding noise stimuli (left panel) for V2 (blue) and V4 (red). The right panel depicts the difference in responses normalized by their sum.

Response dynamics for textures and noise images during the control experiment were consistent with those trends. The differences in responses to the textures and the noise images were more pronounced in V4 than in V2 (Fig. 7*D*). Figure 7*E* left depicts the differences in normalized firing rates between the textures and noise images (Modulation index), which indicates the strength of the modulation by higher-order statistics. The overall modulation in V4 was greater than that in V2 (around 130–300 ms), while the onset of the modulation appears earlier in V2 in this plot. The larger modulation in V4 was also evident when we normalized the modulation to the sum of responses to textures and noise images (i.e. (normalized responses to texture – that to noise)/(normalized responses to texture + that to noise); Fig. 7*E* right). The asynchrony of the onsets between V2 and V4, however, disappeared with this normalized modulation index because V2 exhibited a sharp phasic onset activity (Fig. 7*D*), which increased the denominator of the normalized modulation index and thus reduced the earlier V2 modulation. Together, these analyses of the latency of higher-order components did not enable us to conclude which area took precedence.

Response Characteristics

The aforementioned fitting analyses described the stimulus selectivity but disregarded other response characteristics such as average firing rates and the sparseness of responses. Here we compare those response characteristics between V2 and V4. In particular, sparseness is thought to be indicative of underlying neural mechanisms (Olshausen and Field 1996; Karklin and Lewicki 2009; Carlson et al. 2011), and there has been a debate over whether response sparseness undergoes a systematic change along the cortical hierarchy (Baddeley et al. 1997;

Willmore et al. 2011; Rust and DiCarlo 2012). Rust and DiCarlo (2012) showed that the sparseness of responses to natural images is stable across the ventral pathway, which they explained as reflecting balanced increases in selectivity for images and tolerance to image transformations such as object translation or rotation. Nonetheless, response sparseness for textures may provide different results if the mechanisms underlying the tuning properties differ between objects and textures.

Indeed, we found that V4 neurons responded more sparsely to our textural stimuli than V2 neurons (Fig. 8A). In this test, we used a conventional measure of sparseness (sparseness index; see eq. 2 in Materials and Methods) whose values range from 0 (nonsparse) to 1 (sparse). Neurons in V4 tended to have significantly higher-sparseness indices than those in V2 ($P = 0.0061$, Mann–Whitney U test; mean V2 = 0.42, V4 = 0.52), although it did not reach a significant level in one monkey (monkey SI, $P = 0.18$; monkey EV, $P = 0.05$). Because the sparseness index is sensitive to a multitude of confounding factors (Lehky et al. 2005; Rust and DiCarlo 2012), we performed several control analyses. We first examined a different index of the sharpness of neuronal selectivity, entropy (eq. 3 in Materials and Methods), which is unaffected by the mean and SD of the firing rates (Lehky et al. 2005). Entropy also significantly differed between V2 and V4 (Fig. 8B; $P < 0.001$, Mann–Whitney U test; monkey SI, $P = 0.0033$; monkey EV, $P = 0.071$). Because the sparseness index shows a systematic bias when spikes are generated from Poisson process (Rust and DiCarlo 2012), we adopted the corrected sparseness index proposed in Rust and DiCarlo (2012), which confirmed that the significant difference persisted ($P = 0.013$). In our experiment, we sampled neural responses using adaptive exploration, which could be a potential source of artifact in the estimation of sparseness. We therefore calculated sparseness using only data from the first generation of sampling, which was unaffected by the adaptive exploration, and found similar results (V2 = 0.41; V4 = 0.53; $P = 0.0022$).

We also observed that the firing rates averaged across all cells and stimuli were higher in V2 than in V4, as was the SD of firing rates across stimuli (Fig. 8D). This could not be solely explained by the difference in sparseness since the maximum firing rates for the most preferred texture were also largely different between the areas (V2: 80.0 spike/s, V4: 59.4 spike/s; $P < 0.001$, Mann–Whitney U test). By contrast, the spontaneous firing rates were statistically indistinguishable between the 2 areas ($P = 0.095$). Intriguingly, the discrimination index, an index of neurons' ability to discriminate stimuli (eq. 1 in Materials and Methods), was comparable between the 2 areas (V2 = 0.72; V4 = 0.74; $P = 0.50$, Mann–Whitney U test; Fig. 8C), despite the differences in average firing rates. The discrimination index incorporates maximum and minimum firing rates as well as response variability (eq. 1 in Materials and Methods). If we excluded the variability from the discrimination index (i.e. the index defined as the difference in the maximum and minimum firing rates divided by their sum), V4 outperformed V2 (V2 = 0.82; V4 = 0.92; $P < 0.001$). These results, showing sparser and lower responses in V4, indicate that, overall, V4 tends to respond to textural stimuli more weakly than V2, whereas the discrimination index was indistinguishable or even slightly better in V4.

Discussion

Previous studies reported that surface properties and textures are processed in the ventral visual pathway (Hanazawa and Komatsu 2001; Liu et al. 2004; Köteles et al. 2008; Nishio et al. 2012; Okazawa et al. 2012; Goda et al. 2014; Nishio et al. 2014; Orban et al. 2014), but how responses in V1 are integrated into naturalistic texture representations remains an enigma. Given the recent studies suggesting that both V2 and V4 respond to higher-order image statistics (i.e. combinations of V1-like filter responses) (Freeman et al. 2013; Okazawa et al. 2015; Ziemba

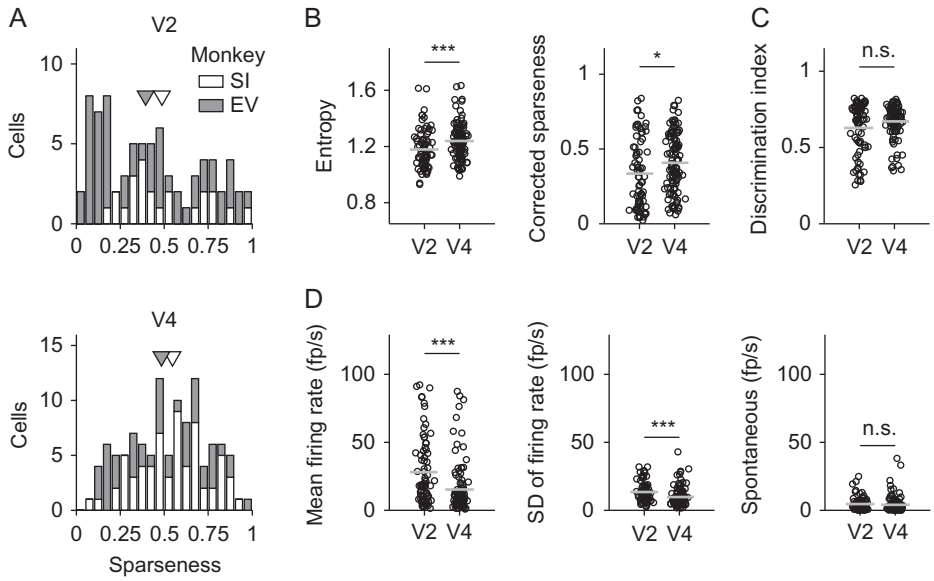


Figure 8. Comparison of firing properties in V2 and V4. The analyses were performed using all the cells recorded (78 cells in V2 and 109 cells in V4). (*A*) Sparseness index, which represents the sharpness of stimulus selectivity (see "Analysis of neuronal firing rates" in Materials and Methods for its definition). The 2 bar brightnesses indicate the different monkeys: white, monkey SI; gray, monkey EV. The upper and lower panels correspond to V2 and V4, respectively. Arrowheads indicate the average values for each monkey. (*B*) Distribution of entropy and sparseness indices corrected for the Poisson properties of the firing rates. Each dot represents one neuron, and their horizontal positions are jittered for visualization purposes. Horizontal gray bars indicate the mean values. *$P < 0.05$, ***$P < 0.001$, Mann–Whitney U test. (*C*) Distribution of discrimination indices, which represent the neurons' abilities to discriminate textures (see eq. 1 in "Analysis of neuronal firing rates" in Materials and Methods). n.s.: nonsignificant, Mann–Whitney U test. (*D*) Mean, SD, and spontaneous firing rates. ***$P < 0.001$, Mann–Whitney U test.

et al. 2016), the present study aimed to clarify differences in texture selectivity between V2 and V4. Our fitting analysis suggested that, although both areas similarly had weights for V1-level spectral and higher-order features, V4 showed slightly stronger weights for higher-order statistics than V2. We confirmed this finding in a control experiment in which we measured responses to noise images that lacked higher-order structures. These results suggest that representation of higher-order statistics, which is likely a cornerstone of visual texture processing (Freeman and Simoncelli 2011), gradually develops along the ventral pathway.

Despite the widely accepted notion that V2 and V4 serve as important building blocks of visual processing in the ventral pathway, surprisingly few attempts have been made to directly compare the 2 regions using parametrically defined stimuli. Hegdé and Van Essen (2007) compared the 2 regions using non-Cartesian (circular or hyperbolic) gratings, but found little evidence for the development of stimulus selectivity between V2 and V4. By showing the parametric tunings for higher-order image statistics, the present study, for the first time, revealed systematic differences in selectivity for complex visual patterns between macaque V2 and V4.

It should be noted, however, that we do not specifically argue that V2 and V4 encode the exact model we developed in this report (i.e. parameters reduced from the texture synthesis model). Although a cross-validation analysis showed that no other texture model outperforms our model in terms of fitting accuracy (Supplementary Figure 2), this does not mean that the our model per se is encoded in V2 and V4. Instead, we expect that slightly different forms of our model or other models with similar underlying computations will achieve similar levels of fitting performance. Nonetheless, we believe that our fitting analyses were sufficient to reveal the major similarities and differences in texture selectivity between V2 and V4; that texture selectivity in both areas can be better explained using higher-order features than using only spectral features; and that this tendency is observed slightly more clearly in V4 than V2.

Excluding Possible Confounding Factors

A multitude of factors can induce apparent differences in stimulus selectivity between 2 brain regions. Therefore, special caution is required when comparing them. These confounding factors include the positions and sizes of receptive fields, the effect of the receptive field surround, the recording histories, and the experimental strategies used, such as the adaptive sampling method we used. Here we consider the potential effects of these factors.

We recorded neurons so that the positions of the receptive field center in V2 and V4 closely matched, but the sizes of the receptive fields differed between the 2 areas. We therefore conducted a post hoc analysis and showed that the stimulus size does not account for the fitting results (see the "Tuning for image statistics" section in the Results). It has been recently reported that larger sizes of texture stimuli yielded stronger sensitivity to higher-order features in V2 (Ziemba et al. 2013), but our stimuli should be large enough (6.4°) to drive higher-order responses of the V2 neurons because their receptive field sizes (2.4°) were on average much smaller than the stimulus size. The mean size of V4 receptive field (6.0°) was similar to the stimulus size. Therefore, unless we assume that stimulating the receptive field surround weakens the selectivity for higher-order image statistics (as opposed to Ziemba et al. 2013), the difference in higher-order selectivity between V2 and V4 cannot be explained by the size dependence of the selectivity.

Because we collected all V2 neurons after the V4 recordings, long-term changes in neuronal properties are a potential confounding factor. To assess this possibility, we ran a post hoc analysis to examine a correlation between the tunings of cells and the recording histories. In neither monkeys nor within any experiment on V2 and V4, did we see consistent changes in the strength of selectivity for higher-order features during the course of the experiments. It should also be noted that recordings from V2 and V4 were made in different hemispheres in both monkeys. Therefore, invasions of the downstream region (V4) are unlikely to account for the weaker selectivity for higher-order features in V2.

Finally, we should consider the effect of our specific strategy of collecting neuronal responses, that is the adaptive sampling method. Although adaptive sampling is advantageous for efficient sampling of neuronal tunings, the biased sampling could affect the fitting results. However, we do not think that this is the case for the following reasons. First, our adaptive sampling procedure always included randomly selected stimuli in each generation (see Materials and Methods), which would mitigate the effect of sampling bias. Second, using a subset of V4 cells, we previously confirmed that the adaptive sampling converged to similar preferred textures when started from 2 different initial sets of textures (Okazawa et al. 2015), which indicates that the procedure yields robust results. Third, our main result, that V4 had greater sensitivity to higher-order features, was also supported by an independent control experiment using noise images.

Roles of V2 and V4 in the Processing of Texture and Other Visual Features

Our results from V2 are largely consistent with earlier studies. Freeman et al. (2013) showed that higher-order features activate V2 neurons more strongly than spectrally matched noise images, which lack those features. The present study found a similar level of modulation by higher-order features in V2, but there was a small difference in the time courses of the modulation; V2 neurons in our study displayed strong modulation almost from the onset of visual responses, whereas neurons in their study showed a gradual increase in modulation during late parts of visual responses (Freeman et al. 2013, Figure 2c). Our PSTH in V2 also exhibited a sharp response onset, which is probably because we used awake animals. Whether the earlier rise in the modulation for higher-order features can also be explained by differences between awake and anesthetized states remains to be determined.

Neurons in both V2 and V4 respond to moderately complex shapes such as curved contours or non-Cartesian gratings (Gallant et al. 1993; Hegdé and Van Essen 2000; Ito and Komatsu 2004; Anzai et al. 2007; Nandy et al. 2013; Yu et al. 2015). As mentioned earlier, Hegdé and Van Essen (2007) found little evidence for the development of selectivity for non-Cartesian gratings from V2 to V4, which may be because the selectivity is attributable to spectral parameters (David et al. 2006). On the other hand, V4 neurons show selectivity for curvatures at specific locations within shapes (Pasupathy and Connor 2001), which cannot be accounted for by the selectivity for spectral parameters (Oleskiw et al. 2014). Because the higher-order texture representation also does not account for object-centered contour representations as the texture representation lacks spatial information, we favor the idea that

representations of texture and contours develop separately along the ventral pathway, although this does not necessarily mean that different populations of neurons represent texture and contours. Advances in hierarchical network models may provide an account for both texture and shape representations (Khaligh-Razavi and Kriegeskorte 2014; Yamins et al. 2014; Gatys et al. 2015; Güçlü and van Gerven 2015).

Interpretation of Differences in Firing Characteristics

In addition to the texture-tuning properties, we found characteristic differences between V2 and V4 in their firing properties. V4 had lower average firing rates and responded more sparsely to textural stimuli. This indicates that textural stimuli, on average, evoked fewer spikes in V4 than in V2 at the single-cell level. This result appears to be consistent with a previous functional magnetic resonance imaging finding (Freeman et al. 2013) that V2 exhibited stronger activation in response to textural stimuli than does V4. This does not, however, readily indicate that V2 plays a more important role in texture processing, since V2 and V4 had almost identical powers to explain humans' discriminability of textures and categorization of materials (see Supplementary Figure 3). Instead, this may suggest that V4 neurons represent textures more economically than V2, considering that V4 encodes a similar level of information using smaller numbers of spikes.

Previous studies have shown that the sparseness of responses to natural scenes is unchanged along the ventral pathway (Willmore et al. 2011; Rust and DiCarlo 2012), which is explained by a constant balance of selectivity for object features and tolerances of object transformations (e.g. spatial shifting) within neuronal receptive fields (Rust and DiCarlo 2012). Compared with natural scenes consisting of objects, our textural stimuli are less concerned with the tolerance of responses due to their repetitive structures. This may explain the observed difference in sparseness between V2 and V4 in our study. Weaker V4 responses to textures may also imply that V4 neurons efficiently allocate their resources for encoding/processing more complex features than the higher-order statistics considered here, such as conjunctions of multiple textures and shapes. To fully describe the neural representation in higher-visual cortices, it will be necessary to characterize neurons' tunings by parametrically manipulating visual stimuli containing such complex features.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Funding

JSPS KAKENHI (Grant Numbers 22135007 and 15H05916) (Grant-in-Aid for Scientific Research on Innovative Areas "Shitsukan" and "Innovative SHITSUKAN Science and Technology" to H.K.); Grant-in-Aid for JSPS Fellows from the Japan Society for the Promotion of Science to G.O; Center of Innovation Program from Japan Science and Technology Agency, JST.

Notes

The authors would like to thank M. Takagi and T. Ota for technical assistance; C. Ziemba for valuable comments on the manuscript. *Conflicts of interest:* None declared..

References

Adelson EH. 2001. On seeing stuff: the perception of materials by humans and machines. In: Rogowitz BE, Pappas TN, editors. Proceedings of the SPIE. Volume 4299: Human Vision and Electronic Imaging VI. Bellingham, WA: SPIE. p. 1–12.

Anzai A, Peng X, Van Essen DC. 2007. Neurons in monkey visual area V2 encode combinations of orientations. Nat Neurosci. 10:1313–1321.

Arcizet F, Jouffrais C, Girard P. 2008. Natural textures classification in area V4 of the macaque monkey. Exp Brain Res. 189:109–120.

Baddeley R, Abbott LF, Booth MC, Sengpiel F, Freeman T, Wakeman EA, Rolls ET. 1997. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proc Biol Sci. 264:1775–1783.

Bishop CM. 2006. Pattern recognition and machine learning. New York: Springer.

Brincat SL, Connor CE. 2006. Dynamic shape synthesis in posterior inferotemporal cortex. Neuron. 49:17–24.

Carlson ET, Rasquinha RJ, Zhang K, Connor CE. 2011. A sparse object coding scheme in area V4. Curr Biol. 21:288–293.

David SV, Hayden BY, Gallant JL. 2006. Spectral receptive field properties explain shape selectivity in area V4. J Neurophysiol. 96:3492–3505.

El-Shamayleh Y, Movshon JA. 2011. Neuronal responses to texture-defined form in macaque visual area V2. J Neurosci. 31:8543–8555.

Fleming RW. 2014. Visual perception of materials and their properties. Vision Res. 94:62–75.

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. 2013. A functional and perceptual signature of the second visual area in primates. Nat Neurosci. 16:974–981.

Gallant JL, Braun J, Van Essen DC. 1993. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science. 259:100–103.

Gatys LA, Ecker AS, Bethge M. 2015. Texture synthesis using convolutional neural networks. In: Advances in Neural Information Processing Systems 28 (NIPS 2015). p. 262–270.

Goda N, Tachibana A, Okazawa G, Komatsu H. 2014. Representation of the material properties of objects in the visual cortex of nonhuman primates. J Neurosci. 34:2660–2673.

Güçlü U, van Gerven MA. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci. 35:10005–10014.

Hanazawa A, Komatsu H. 2001. Influence of the direction of elemental luminance gradients on the responses of V4 cells to textured surfaces. J Neurosci. 21:4490–4497.

Hegdé J. 2008. Time course of visual perception: coarse-to-fine processing and beyond. Prog Neurobiol. 84:405–439.

Hegdé J, Van Essen DC. 2000. Selectivity for complex shapes in primate visual area V2. J Neurosci. 20:61–66.

Hegdé J, Van Essen DC. 2004. Temporal dynamics of shape analysis in macaque visual area V2. J Neurophysiol. 92:3030–3042.

Hegdé J, Van Essen DC. 2007. A comparative study of shape representation in macaque visual areas V2 and V4. Cereb Cortex. 17:1100–1116.

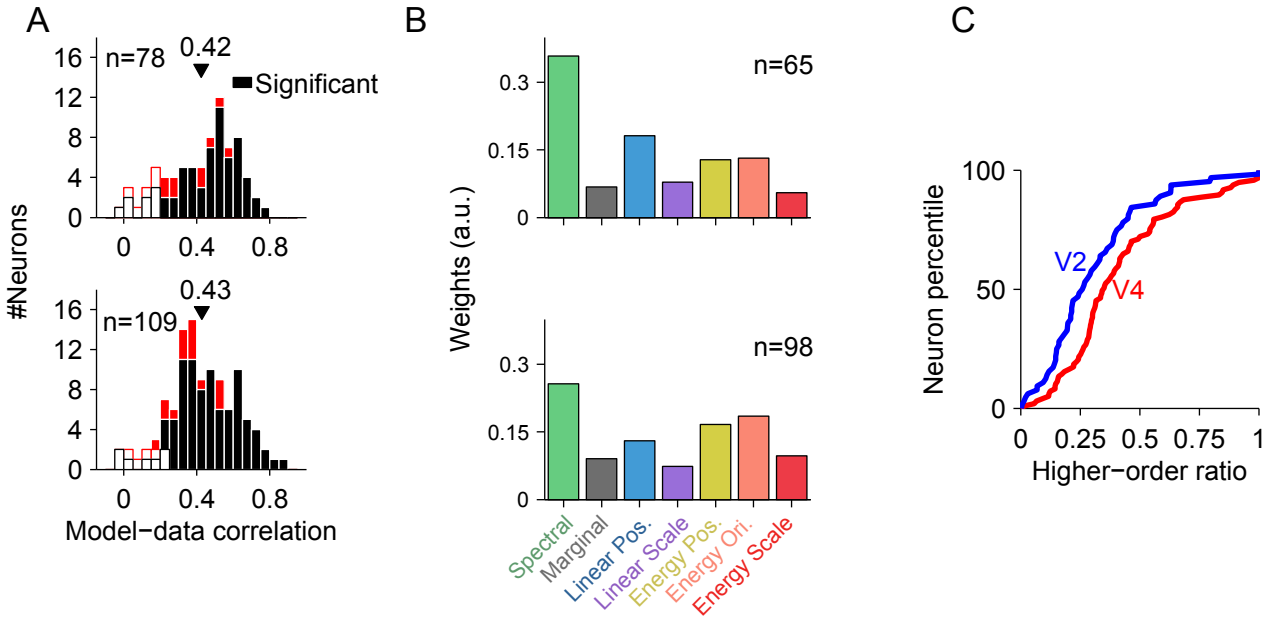
Hiramatsu C, Goda N, Komatsu H. 2011. Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. NeuroImage. 57:482–494.

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

Karklin Y, Lewicki MS. 2009. Emergence of complex cell properties by learning to generalize in natural scenes. Nature. 457:83–86.
Khaligh-Razavi S-M, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 10:e1003915.
Kohler PJ, Clarke A, Yakovleva A, Liu Y, Norcia AM. 2016. Representation of maximally regular textures in human visual cortex. J Neurosci. 36:714–729.
Köteles K, De Maziere PA, Van Hulle M, Orban GA, Vogels R. 2008. Coding of images of materials by macaque inferior temporal cortical neurons. Eur J Neurosci. 27:466–482.
Lehky SR, Sejnowski TJ, Desimone R. 2005. Selectivity and sparseness in the responses of striate complex cells. Vision Res. 45:57–73.
Liu Y, Vogels R, Orban GA. 2004. Convergence of depth from texture and depth from disparity in macaque inferior temporal cortex. J Neurosci. 24:3795–3800.
Nandy AS, Sharpee TO, Reynolds JH, Mitchell JF. 2013. The fine structure of shape tuning in area V4. Neuron. 78:1102–1115.
Nishio A, Goda N, Komatsu H. 2012. Neural selectivity and representation of gloss in the monkey inferior temporal cortex. J Neurosci. 32:10780–10793.
Nishio A, Shimokawa T, Goda N, Komatsu H. 2014. Perceptual gloss parameters are encoded by population responses in the monkey inferior temporal cortex. J Neurosci. 34: 11143–11151.
Okazawa G, Goda N, Komatsu H. 2012. Selective responses to specular surfaces in the macaque visual cortex revealed by fMRI. Neuroimage. 63:1321–1333.
Okazawa G, Tajima S, Komatsu H. 2015. Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc Natl Acad Sci USA. 112:E351–360.
Oleskiw TD, Pasupathy A, Bair W. 2014. Spectral receptive fields do not explain tuning for boundary curvature in V4. J Neurophysiol. 112:2114–2122.
Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 381:607–609.
Orban GA, Qi Z, Vanduffel W. 2014. The transition in the ventral stream from features to real world entities representations. Front Psychol. 5:695.
Pasupathy A, Connor CE. 2001. Shape representation in area V4: position-specific tuning for boundary conformation. J Neurophysiol. 86:2505–2519.
Portilla J, Simoncelli EP. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis. 40:49–71.
Prince SJ, Pointon AD, Cumming BG, Parker AJ. 2002. Quantitative analysis of the responses of V1 neurons to horizontal disparity in dynamic random-dot stereograms. J Neurophysiol. 87:191–208.
Rust NC, DiCarlo JJ. 2012. Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. J Neurosci. 32:10170–10182.
Sanada TM, Nguyenkim JD, Deangelis GC. 2012. Representation of 3-D surface orientation by velocity and disparity gradient cues in area MT. J Neurophysiol. 107:2109–2122.
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J R Stat Soc B. 58:267–288.
Vinje W, Gallant J. 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. Science. 287: 1273–1276.
Willmore BD, Mazer JA, Gallant JL. 2011. Sparse coding in striate and extrastriate visual cortex. J Neurophysiol. 105: 2907–2919.
Yamane Y, Carlson ET, Bowman KC, Connor CE. 2008. A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nat Neurosci. 11:1352–1360.
Yamins DLK, 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.
Yu Y, Schmid AM, Victor JD. 2015. Visual processing of informative multipoint correlations arises primarily in V2. eLife. 4: e06604.
Ziemba CM, Freeman J, Movshon JA, Simoncelli EP. 2016. Selectivity and tolerance for visual texture in macaque V2. Proc Natl Acad Sci USA. 133:E3140–E3149.
Ziemba CM, Freeman J, Simoncelli EP, Movshon JA. 2013. Size dependence of sensitivity to naturalistic stimuli in macaque V2. In: Society for Neuroscience Annual Meeting. San Diego, California.

Supplementary information

Figure S1



Fitting results without excluding sparse neurons. In the main fitting analyses (Fig. 5), we excluded a portion of the cells (V2: 14 cells, V4: 19 cells) whose sparseness index was greater than 0.75. This is because the fitting could be unreliable when a neuron exhibits responses to only a small number of textures, as the information gained from a small number of stimuli is limited. To confirm that this selection procedure did not introduce bias into the results, we performed the same fitting analyses with those cells included.

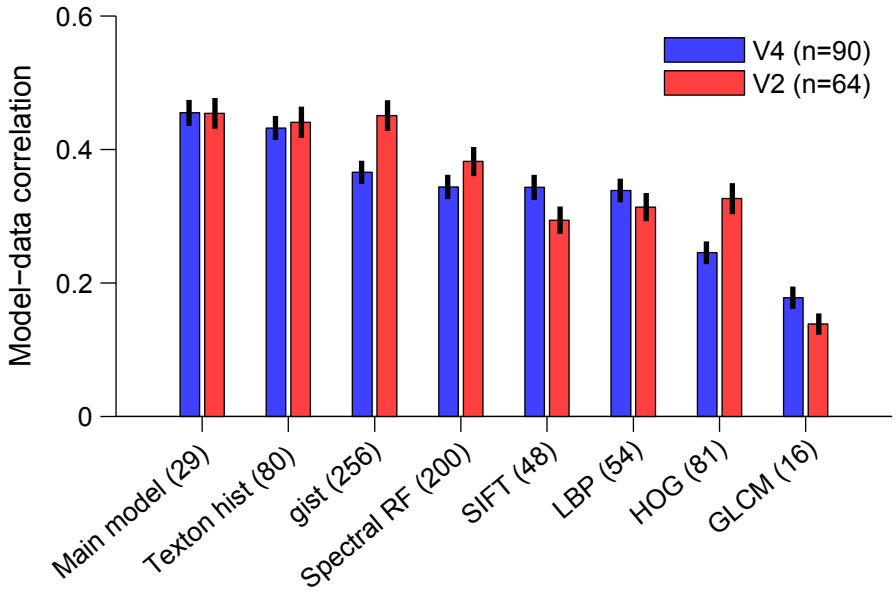
(A) Correlations between the observed and model-predicted firing rates, determined using cross-validation (corresponding to Fig. 5A). Red bars indicate cells excluded in the main analyses. Filled bars represent cells that could be fitted significantly better than chance. The mean correlation between the observed and predicted firing rates were 0.42 and 0.43 for V2 and V4, respectively.

(B) Amplitudes of fitting weights averaged across neurons, which corresponds to Fig. 5C.

(C) Cumulative histogram of higher-order ratios, which corresponds to Fig. 5E. There was a significant difference in higher-order ratios between V2 and V4 (p = 0.0029; Mann-Whitney U test).

Figure S2

A



B

	Non local	Spectral	Higher order
Main model	○	○	○
Texton hist	○	○	○
gist	△	○	×
Spectral RF	○	○	×
SIFT	○	×	-
LBP	○	×	-
HOG	×	×	-
GLCM	○	×	-

Comparison of fitting performance achieved with the texture model in the main text ("Main model") and other models. These include models describing textures (Texton hist, LBP, GLCM) or objects or scenes (gist, SIFT, HOG) and those previously used to describe neuronal activities in V4 (Spectral RF) (David et al., 2006). Brief explanations and implementations of the models are provided below. Numbers in the parenthesis indicate the number of parameters in each model. The error bars indicate the S.E.M. across the neurons. The V4 data were adopted from Okazawa et al. (2015).

Using each model, we computed the statistical parameters of the texture images and fitted those parameters to the neurons' responses, as we did with our texture model in the main text. In brief, we performed regularized (L1) linear regression (Tibshirani, 1996) to obtain the fitting weights using 90% of the stimuli. We used the remaining 10% of stimuli to estimate the fitting performance by computing a Pearson's correlation coefficient between the observed and predicted firing rates (y-axis in Fig. S2A). Because the fitting performance was tested using those independent 10% test sets, the result is expected to be largely immune to differences in the numbers of parameters across models.

We also described several properties of the models based on the underlying computations performed in each model (Fig. S2B). "Non local" means that features computed in a given model do not depend on the particular spatial positions in an image. The triangle in "Non-local" indicates that features are partially localized. "Spectral" means that a model computes features related to spatial frequency components such as Gabor-like filters. "Higher order" means that a model also

incorporates features related to the correlations among different spatial frequency components.

For both V2 and V4, no model outperformed our main model, although some models achieved nearly identical performances (Fig. S2A). For V4, our main model was statistically better than any of the other models (with "Texton hist": $p = 0.001$, with others: $p < 0.001$; Wilcoxon signed rank test). For V2, our main model was better than all models except "gist" (with "Texton hist": $p = 0.038$, with "gist": $p = 0.59$, with others: $p < 0.001$; Wilcoxon signed rank test). Significant differences between V2 and V4 were observed for "gist" ($p = 0.003$) and "HOG" ($p = 0.006$; Mann-Whitney U test). Overall, the fitting performances across models were similar between V2 and V4 ($r = 0.88$, $p = 0.004$).

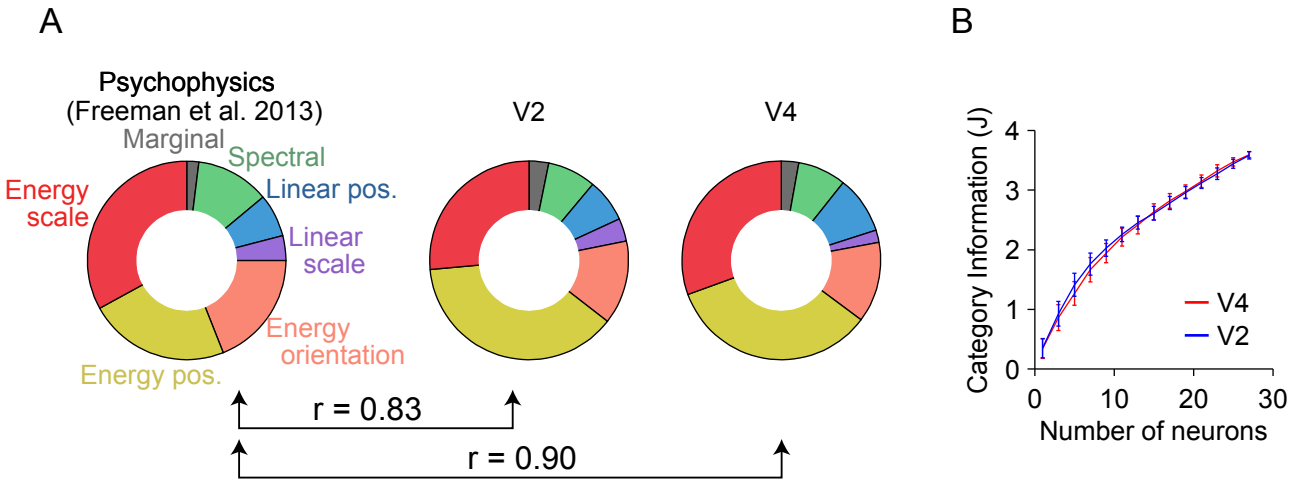
We found several consistencies between the performances of models and model properties (compare S2A and S2B). For V4, models using higher-order features tended to perform better. For V2, models using spectral features tended to perform well. Although these analyses remain anecdotal because the results may depend on the choice of models and model parameters, the overall results support our claim that both V2 and V4 have similar tuning properties and that V4 has slightly stronger selectivity for higher-order parameters.

Note that all of these models explicitly compute the statistical properties of images. Although a recent study showed that hierarchical network models achieve good fitting performance to the neural responses in V4 (Yamins et al., 2014), we did not include them because those models do not explicitly represent the statistical parameters to compute, and the image features that evoked the neuronal responses are not known.

Implementation of the models: We basically implemented the models and determined the model parameters essentially as described in the original papers describing these models. We implemented or made use of available codes of many existing models and computed their statistical parameters for all presented textures. Because each model has several variants and some hyperparameters, we will briefly describe how we implemented the models. 'Texton histogram' (Varma and Zisserman, 2005) is the model that describes the co-occurrence of different spatial-frequency/orientation subbands as a texton and counts up the numbers of occurrences of individual textons in an image. The number of textons is set 80. 'Gist' (Oliva and Torralba, 2001) is a model that concatenates the spatial frequency components of subregions of an image. We split an image into 4 by 4 subregions. 'Spectral RF' (David et al., 2006) is introduced to explain the V4 responses to spatial frequency components. We computed the amplitudes of Fourier transforms in a range of spatial frequencies from the DC

component to the 10 cycles/image components. Because we intended to compute the fitting performance, we did not take into account the intrinsic correlations between pairs of spectral channels in the stimuli, as was done in the paper (David et al., 2006). ‘SIFT’ (Lowe, 2004) is the model that captures the local gradient pattern. To do so, the model first finds several keypoints in an image and computes the local orientations of gradients around each keypoint. To describe an image, the model counts up the numbers of occurrences of several dominant local patterns. The number of dominant local patterns is up to the users, and we set it as 48. Similarly, ‘LBP’ (Ojala et al., 2002) counts up the numbers of occurrences of several dominant local patterns, but the way of representing local patterns is different. ‘HOG’ (Dalal and Triggs, 2005; Ludwig et al., 2009) computes the local gradients in subregions of an image. We split an image into 3 by 3 sub-regions. ‘GLCM’ (Haralick et al., 1973) first computes the autocorrelations of pixels and extracts several features describing the auto-correlations. We specifically extracted features called ‘Contrast’, ‘Energy’, ‘Homogeneity’ and ‘Entropy’.

Figure S3



Comparison of neuronal selectivity with perceptual discriminability of textures and separability of material categories. The V4 data was adopted from Okazawa et al. (2015).

(A) Close correspondence between neuronal selectivity and perceptual discriminability. The left panel, borrowed from Fig. 7d in Freeman et al. (2013), shows the relative powers of each group of statistics to explain humans' ability to discriminate textures from spectrally-matched noise images. For the neural data, we estimated differences in responses to textures and noise images for V2 and V4 cells based on the fitting weights. We then computed the relative contributions of each group of statistics for these differential responses (middle and right panels, respectively; see Methods below for the details). Pearson's correlation coefficients between the psychophysical result and V2 (V4) were 0.83 (0.90) (V2: $p = 0.019$, V4: $p = 0.006$; test for significance of Pearson's correlation coefficient). The difference in the correlation coefficients between the areas was marginally significant ($p = 0.055$, permutation test).

(B) Abilities of neurons in V2 (red) and V4 (blue) to separate different material categories. We randomly chose n neurons (abscissa) and simulated how well textures of different material categories are separated by responses of those n neurons (see methods below). We observed very similar category separability in V2 and V4 (Fig. 7B), which may seem at odd with the stronger selectivity for higher-order statistics in V4. We suspect that it comes from the fact that spectral features can also contribute to the categorization of material properties (Giesel and Zaidi, 2013), which could improve the performance of V2.

Methods:

[Texture discriminability] We computed the texture discriminability of neurons in a manner

comparable to the way Freeman et al. (2013) computed humans' texture discriminability. They psychophysically measured the threshold of humans' ability to discriminate between naturalistic textures and noise images generated by randomizing the phase of the spatial frequency components of the textures in the Fourier space. To perform a neuronal analysis comparable to the psychophysical one, we first computed all neurons' predicted firing rates for our 10355 textures and their corresponding noise images. The predicted firing rates were computed using the neurons' linear fitting weights obtained in the fitting analysis. We then estimated the neurons' population discriminability of each texture using the following equation:

$$d_t = \sqrt{\sum_i \frac{(FRT_i - FRN_i)^2}{\sigma_i^2}},$$

where FRT_i and FRN_i are the predicted firing rates of neuron i for a texture and its corresponding noise image, and σ_i^2 is the variance of the i-th neuron's responses obtained by computing the trial-by-trial variances of the firing rates of that neuron. We linearly regressed the obtained neural discriminability (d_t) for all textures using the spectral and higher-order statistical parameters of the images (Fig. 1C) to determine which image parameters contribute to the discriminability. For example, if neurons are better at discriminating textures having a particular parameter, the discriminability (d_t) of those neurons should be well fitted by that parameter. We computed the relative contribution of each group of parameters for this linear regression using a method called averaging-over-orderings (Grömping, 2007). This method computes a difference in the R^2 of the regression before and after inclusion of a particular group of parameters. That difference corresponds to the size of the contribution of that group, thus yielding the average percent contributions of individual groups to explain the neural discriminability of textures. We quantified the similarity of the contributing statistical groups between the psychophysical data (Freeman et al., 2013) and our neuronal data using Pearson's correlation coefficient of percent contributions across groups. We performed a permutation test to statistically examine the difference in these correlation coefficients between V2 and V4. We shuffled neurons between V2 and V4 and computed the difference in the correlation coefficient for the shuffled data. This procedure was repeated 2000 times to test whether the actual difference in the correlation coefficient is significantly larger than the difference in the shuffled data.

[Separability of material categories] To calculate the separability of material categories, we used 4170 texture stimuli in our stimulus set, which were tagged with eight different material categories (bark, sand, fabric, fur, leather, stone, water, and wood; average of 521 images in each). We

randomly selected n neurons and computed the predicted firing rates for those textures based on the linear fitting weights obtained in the fitting analysis. Thus, the 4170 images were projected into an n-dimensional space where dimensions correspond to neurons. We formulated category separability (J) in this n-dimensional space using the following equation:

$$J(\mathbf{w}) = \ tr\left[\frac{\boldsymbol{wS_Bw^T}}{\boldsymbol{wS_Ww^T}}\right],$$

where S_B is the between-category covariance matrix and S_W is the within-category covariance matrix of stimuli. We randomly chose n neurons 100 times to estimate the average J as a function of the number of neurons (n). A larger J indicates greater separability of material categories. We used the category separability instead of classification accuracies because it does not depend on specific algorithms of multi-class classifications.

SI References

1. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition 1:886-893.
2. David SV, Hayden BY, Gallant JL (2006) Spectral receptive field properties explain shape selectivity in area V4. J Neurophysiol 96:3492-3505.
3. Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA (2013) A functional and perceptual signature of the second visual area in primates. Nat Neurosci 16:974-981.
4. Giesel M, Zaidi Q (2013) Frequency-based heuristics for material perception. J Vis 13.14:7.
5. Grömping U (2007) Estimators of relative importance in linear regression based on variance decomposition. The American Statistician 61:139-147.
6. Haralick RM, Shanmuga.K, Dinstein I (1973) Textural features for image classification. IEEE Transactions on Systems Man and Cybernetics SMC-3:610-621.
7. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60:91-110.
8. Ludwig O, Delgado D, Goncalves V, Nunes U (2009) Trainable classifier-fusion schemes: An application to pedestrian detection. In: 12th International IEEE Conference On Intelligent Transportation Systems, pp 1-6. St. Louis.
9. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:971-987.
10. 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:E351-360.
11. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42:145-175.
12. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological):267-288.
13. Varma M, Zisserman A (2005) A statistical approach to texture classification from single images. Int. J. Comput. Vision 62:61-81.
14. Yamins DLK, 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 U S A 111:8619-8624.