Comparing visual representations across human fMRI and computational vision

Daniel D. Leeds
Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA
Department of Computer and Information Science, Fordham University, Bronx, NY, USA

Darren A. Seibert
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

John A. Pyles
Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Michael J. Tarr
Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Feedforward visual object perception recruits a cortical network that is assumed to be hierarchical, progressing from basic visual features to complete object representations. However, the nature of the intermediate features related to this transformation remains poorly understood. Here, we explore how well different computer vision recognition models account for neural object encoding across the human cortical visual pathway as measured using fMRI. These neural data, collected during the viewing of 60 images of real-world objects, were analyzed with a searchlight procedure as in Kriegeskorte, Goebel, and Bandettini (2006): Within each searchlight sphere, the obtained patterns of neural activity for all 60 objects were compared to model responses for each computer recognition algorithm using representational dissimilarity analysis (Kriegeskorte et al., 2008). Although each of the computer vision methods significantly accounted for some of the neural data, among the different models, the scale invariant feature transform (Lowe, 2004), encoding local visual properties gathered from "interest points," was best able to accurately and consistently account for stimulus representations within the ventral pathway. More generally, when present, significance was observed in regions of the ventral-temporal cortex associated with intermediate-level object perception. Differences in model effectiveness and the neural location of significant matches may be attributable to the fact that each model implements a different featural basis for representing objects (e.g., more holistic or more parts-based). Overall, we conclude that well-known computer vision recognition systems may serve as viable proxies for theories of intermediate visual object representation.

Introduction

The process of visual object recognition typically associates visual inputs—commencing with an array of light intensities falling on the retina—with semantic categories, for example, "cow," "car," or "face." Nearly every model, theory, or computational system that attempts to implement or account for this process, including the biological visual recognition system realized in the ventral occipito-temporal pathway of the human brain, assumes a feedforward visual processing hierarchy in which the features of representation progressively increase in complexity as one moves up in a feedforward manner (Riesenhuber & Poggio, 1999)—the ultimate output being high-level *object representa-*

Citation: Leeds, D. D., Seibert, D. A., Pyles, J. A., & Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *Journal of Vision*, *13*(13):25, 1–27, http://www.journalofvision.org/content/13/13/25, doi:10.1167/13.13.25.

doi: 10.1167/13.13.25 Received March 18, 2013; published November 22, 2013 ISSN 1534-7362

tions that allow the assignment of category-level labels. It goes almost without saying that within this framework one presupposes levels of *intermediate* featural representations that, while less complex than entire objects, nonetheless capture important object-level visual properties (Ullman, Vidal-Naquet, & Sali, 2002). Yet, for all the interest in uncovering the nature of such features with respect to biological vision, they remain remarkably elusive. At present, there is little empirical data on the neural representations of visual objects in the netherworld between input image and object representation. The goal of our present study is to unravel how the human brain encodes object information along the ventral pathway—the neural "real estate" associated with visual object processing.

Given the paucity of data that bears on this question, how do we develop viable theories explicating the features underlying the neural representation of objects? One possibility is to focus on feature codes realized in "category-selective" regions within the ventral-temporal cortex. However, most investigations of these regions—for example, the "fusiform face area" (FFA) associated with the detection and discrimination of faces (Haxby, Hoffman, & Gobbini, 2000; Grill-Spector, Knouf, & Kanwisher, 2004), the "parahippocampal place area" (PPA) associated with scene processing (Epstein, Harris, Stanley, & Kanwisher, 1999), or the lateral occipital complex (LOC) associated with the processing of objects more generally (Grill-Spector, Kourtzi, & Kanwisher, 2001)—emphasize specific object-level experiential factors or input characteristics that lead to their recruitment but never establish the underlying visual properties that form the basis of the nominally category-specific representations. Most studies of the visual properties that lead to the recruitment of these class-specific, functionally defined brain regions have focused on the effects of spatial transformations and of the alteration of simple domain-specific features (Tsao & Livingstone, 2008). For example, images of objects from within a given class often elicit similar neural responses when scaled, rotated, or moved to different locations in the visual field although, in the case of picture-plane inversion or 3-D rotation, there is typically some change in neural activity (Perrett et al., 1984; Haxby et al., 1999). To the extent that viable models of neural representation have been developed, they have relied on the statistical analysis of the input space within a restricted object domain. For example, "face spaces," nominally capturing the featural dimensions of human face representation, can be defined using principal component analysis (PCA) on face images or using parameterized models that are generative for constructing what appear to be realistic new face stimuli (Calder & Young, 2005; Freiwald, Tsao, & Livingstone, 2009). Alternatively, the featural dimensions of representation are sometimes made more explicit as in Kravitz, Peng, and Baker (2011), who found that the encoding of scenes in the human visual cortex can be understood in terms of an underlying set of intuitive properties, including "open/closed" and "natural/artificial."

This is not to say that studies of intermediate feature representation have not provided some more fine-grained data regarding the neural encoding of objects. For example, Tanaka (2003) explored the minimal visual stimulus that was sufficient to drive a given neuron at a level equivalent to the complete object. He found that individual neurons in IT were selective for a wide variety of simple patterns and shapes that bore some resemblance to the objects initially used to elicit a response from each neuron. Interestingly, Tanaka hypothesized that this pattern-specific selectivity is organized into a columnar structure that maps out a high-dimensional feature space for representing visual objects. Similarly, Yamane, Carlson, Bowman, Wang, and Connor (2008) and Hung, Carlson, and Connor (2012) used a somewhat different search procedure employing a highly constrained, parameterized stimulus space to identify contour selectivity for individual neurons in the primate visual cortex. They found that most contour-selective neurons in V4 and IT each encoded some subset of the parameter space. Moreover, each 2-D contour within this space appeared to encode specific 3-D surface properties, and small collections of these contour-selective units were sufficient to capture the overall 3-D appearance of an object or object part. Within the human neuroscience literature, the study most often associated with feature decoding is that of Kay, Naselaris, Prenger, and Gallant (2008), who explored how neural units coding for orientation and scale within human V1, V2, and V3 can be assembled to reconstruct complex images. Although Kay et al. provide an elegant demonstration of how modern fMRI methods may support more fine-grained analyses (and therefore inspiration for further investigation), their work does not inform us regarding the nature of *intermediate features* in that they relied on well-established theories regarding the featural properties of V1 and V2. That is, they decoded features within a reasonably well-understood parameter space in which it is generally agreed that the particular brain regions in question encode information about the orientations and scales of local edges. Indeed, we see this as the fundamental problem in any attempt to decode the features of "intermediate-level" object representation: The parameter space is extremely large and highly underspecified; therefore, it is difficult to find effective models that fit the data. As such, Ullman et al.'s (2002) proposal that intermediate features can be construed as image fragments of varying scale and location—leaving the content of said fragments entirely unspecified—is perhaps the strongest attempt yet at

capturing task-relevant object information encoded within the human ventral pathway. Supporting the effectiveness of this sort of approach, there is some neurophysiological evidence consistent with the fragment framework laid out by Harel, Ullman, Epstein, and Bentin (2007).

Finally, we note that current computational models commonly applied to biological object recognition tend to make only weak assumptions regarding the nature of intermediate features—the exception being Hummel and Biederman (1992), who made very strong assumptions as to the core features used in object representation—unfortunately, in this model, such strong assumptions work against any generality for the model. For example, almost all models employ variants of Gabor filter banks, detecting local edges in visual stimuli to explain selectivities in the primary visual cortex (V1) (Hubel & Wiesel, 1968). Extending this approach, both Serre, Oliva, and Poggio (2007) and Kay et al. (2008) propose hierarchies of linear and nonlinear spatial pooling computations with Gabor filters at the base to model higher-level vision. One such hierarchical model, HMAX (Cadieu et al., 2007), partially predicts neural selectivity in the midlevel ventral stream (V4) for simple synthetic stimuli. However, HMAX imperfectly clusters pictures of real-world objects relative to clustering based on neurophysiological and fMRI data from IT (Kriegeskorte et al., 2008).

To further address the question of the complex features underlying neural object representation, we employed several models of visual representation drawn from machine vision; each provides a putative hypothesis regarding the features used in object perception. These representations incorporate diverse linear and nonlinear operations on image properties to maximize machine performance in object detection and recognition tasks. As such, we are relying on these models as proxies for theories of features for biological object representation. Given this set of models, we collected data on human object processing using fMRI and a simple object perception task. We then correlated the resultant neural data with the object dissimilarity matrices predicted by each computer vision model, thereby establishing a correspondence between each model and patterns of neural activity in specific spatial locations within the brain. Consistent with the fact that these models make different assumptions with respect to object representation, we found that different models were associated with neural object encoding in different cortical locations. However, consistent with the overall visual nature of all of these representations, we observed that most of these associations lay within the ventral and dorsal visual cortices. Of particular interest, one popular machine vision representation, the scale invariant feature transform (SIFT) (Lowe, 2004), which encodes images using relatively simple local features, was the most strongly associated with measured neural activity in the brain regions typically associated with midlevel object perception (e.g., fusiform cortex). To better explicate how we arrived at this finding, we next define what is meant by "dissimilarity" with respect to both computational models and neural data.

Representational dissimilarity analysis

To assess model performance, neural stimulus representations as measured by fMRI and a given machine vision model were compared using representational dissimilarity analysis. For each set of voxels and for each model, a pairwise distance matrix was computed, reflecting which sets of stimulus images were considered to be similar and which were considered to be different (more detail is given in the Representational Dissimilarity Measures section). Model/neural matrices were more correlated when the two corresponding representations of the stimuli grouped the considered images in a similar manner. Kriegeskorte et al. (2008) demonstrated the advantages of dissimilarity analysis in observing and understanding complex patterns of neural activity—in their case, a collection of spatially contiguous voxels. We similarly wished to understand object encoding across restricted volumes of voxels. The advantage of this approach is that it allows us to judge a model's descriptive power without requiring identification of the exact—most likely nonlinear—mapping between model and voxel responses. Indeed, O'Toole, Jiang, Abdi, and Haxby (2005) and Kiani, Esteky, Mirpour, and Tanaka (2007) pursued related cortical-computational dissimilarity analyses in studying visual perception, finding that the organization of object categories in IT is based, in part, on visual similarity and, in part, on higher-order class information. The ability of this method to bypass the issue of learning direct mapping between model predictions and neural data provides particular benefit for fMRI studies in that it obviates the need to split rather limited data sets in order to cross-validate.

Methods

Stimuli

A picture and word set comprised of 60 distinct color object photos displayed on 53% gray backgrounds and their corresponding basic-level names was used as stimuli (Figure 1). The specific category of each object was selected to match the 60 objects used in Just,

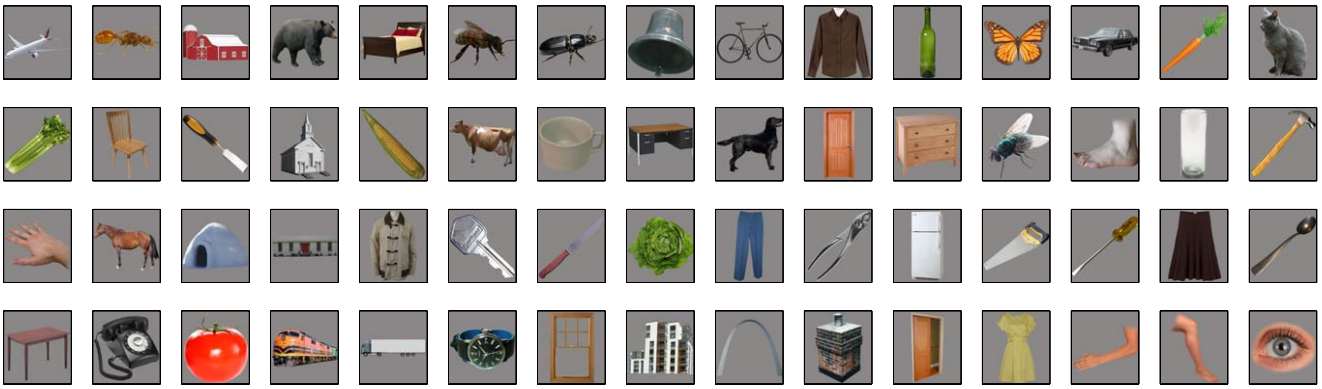

Figure 1. The 60-image stimuli displayed to subjects.

Cherkassky, Aryal, and Mitchell (2010); the particular images used in Just et al. were drawn from the Snodgrass and Vanderwart (1980) line-drawing image data set. The photographic images used in our study were taken from web image searches; therefore, we do not have the rights to redistribute the actual images. The 60 objects included five examples from each of 12 diverse semantic classes, for example, tools, food, mammals, or body parts. Each object was depicted by a single image. Although visual similarities among stimuli can be seen across semantic groups, such as knife and carrot (thin and slanted up to the right) or tomato and eye (circular in the image plane), objects within a semantic class were typically more similar to one another relative to their across-class similarities. Our use of real-world images of objects rather than the hand-drawn or computer-synthesized stimuli employed in the previously discussed studies of midlevel visual coding, for example, Cadieu et al. (2007) and Yamane et al. (2008), is intended to more accurately capture the importance of the broad set of naturally occuring visual features in object perception.

Subjects

Five subjects (one left-handed, one female, age range 20 to 24) from the Carnegie Mellon University community participated, gave written informed consent, and were monitarily compensated for their participation. All procedures were approved by the Institutional Review Board of Carnegie Mellon University.

Experimental design

All stimuli were presented using MATLAB (2012) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) controlled by an Apple Macintosh and were back-projected onto a white screen located at the head end of the bore using a DLP projector (Sharp XG-P560W). Subjects viewed the images through a mirror attached to the head coil with object stimuli subtending a visual angle of approximately 8.3° × 8.3°. Each stimulus was displayed in the center of the screen for 2 s followed by a blank 53% gray screen shown for a time period randomly selected to be between 500 and 3000 ms, followed by a centered fixation cross that remained displayed until the end of each 10-s trial, at which point the next trial began. As such, the stimulus onset asynchrony (SOA) between consecutive stimulus displays was fixed at 10 s. Subjects were instructed to press a button when the fixation cross appeared. The fixation onset detection task was used to engage subject attention throughout the experiment. No other task was required of the subjects, meaning that our study addresses object perception under passive viewing conditions.

The 10 s SOA was chosen to minimize temporal overlap between voxel blood-oxygen level–dependent responses for multiple stimuli: a slow event-related design based on the assumption that the hemodynamic response in the ventral-temporal cortex has decreased to a sufficient degree in the 10–12 s after stimulus onset to minimize the noise in our measurements of the cortical responses.

The stimuli were presented in 24 6-min runs, spread across three 1-hr scanning sessions and arranged to minimize potential adaptation and priming effects. Each scanning session included two sets of four runs. Each run contained 15 word and 15 picture stimuli, ordered such that the picture and the word corresponding to the same object were not viewed in direct succession, and all stimuli were viewed exactly once in each four-run set to avoid priming and adaptation effects. Trials using the word stimuli were not analyzed or otherwise considered as part of our present study. Stimulus order was randomized across blocks and across subjects. Over the course of the experiment, each subject viewed each picture and each word six times; averaging across multiple repetitions was performed for

each stimulus, described below, to reduce trial-by-trial noise.

The first session for each subject also included functional localizer scans to identify object selective cortex, namely, the LOC, a functionally defined region (Kourtzi & Kanwisher, 2000) that we consider separately from the anatomically identified lateral occipital cortex (LO) (although there is overlap between the two areas). For this localizer, 16-s blocks of common everyday objects were alternated with 16-s blocks of phase-scrambled versions of the same objects, separated by 6 s of fixation (Kourtzi & Kanwisher, 2000; Grill-Spector et al., 2001). Phase scrambling was achieved by taking the Fourier transform of each image, randomizing the resulting phase values while retaining the original frequency amplitudes, and reconstructing the image from the modified Fourier coefficients (Schultz & Pilz, 2009). Within each block, 16 images, depicting 14 distinct objects, were shown for 800 ms each, and each object was followed by a 200 ms gray screen. Two of the objects were sequentially repeated once during each block; to maintain attention, subjects were instructed to monitor for this, performing a one-back identity task in which they responded via a key press whenever the same object image was repeated across two image presentations. Six blocks of both the intact and scrambled-object conditions were presented over the 282 s scan (Pyles & Grossman, 2009). The object images used in the localizer scans were different from the object picture stimuli discussed in the Stimuli section. LOC area(s) were identified as those brain regions more selective for intact versus scrambled objects. LOC areas included all regions containing spatially contiguous voxels (no minimum cluster size) for which beta weights for the block design had a significance level of $p < 0.005$.

To provide anatomical information, a T1-weighted structural MRI was performed between runs within the first scanning session for each subject.

fMRI procedures

Subjects were scanned using a 3.0 T Siemens Verio MRI scanner with a 32-channel head coil. Functional images were acquired with a gradient echo-planar imaging pulse sequence (TR 2 s, TE 26 ms, flip angle 90°, $2 \times 2 \times 3$ mm voxels, field of view 192×192 mm^2, 31 oblique-axial slices). Slices spanned the majority of the brain to observe relevant stimulus representations beyond the visual streams (Figure 2). An MP-RAGE sequence (flip angle 9°, 1 mm^3 voxels, field of view 256×256 mm^3, 176 sagittal slices) was used for anatomical imaging.

Preprocessing

Functional scans were coregistered to the anatomical image and motion-corrected using AFNI (Pittman, 2011). Highpass filtering was implemented in AFNI by removing sinusoidal trends with periods of half and full length of each run (338 s) as well as polynomial trends of orders one through three. The data then were normalized so that each voxel's time course was 0 mean and unit variance (Just et al., 2010). To allow multivariate analysis to exploit information present at high spatial frequencies, no spatial smoothing was performed (Swisher et al., 2010). Prior work has demonstrated that spatial smoothing monotonically decreases statistical power for a variety of statistical tests, including univariate t values (Kriegeskorte et al., 2006).

For each stimulus repetition, the measured response of each voxel consisted of five data samples starting 2 s/1 TR after onset, corresponding to the 10 s between stimuli. Each five-sample response was consolidated into a weighted sum, intended to estimate the peak response. This was accomplished as one step in a "searchlight" process (Kriegeskorte et al., 2006): 123-voxel searchlight spheres with radii of 3 voxels were defined centered sequentially on every voxel in the brain. The average five-sample response of voxels across this sphere and across all stimulus presentations was computed. For a given searchlight, for each stimulus, each voxel was assigned a number based on the dot product of this average response and the voxel's mean response across all six repetitions of that stimulus. To the extent that hemodynamic responses are known to vary across cortical regions, this procedure allowed us to take into account a given voxel's local neighborhood mean response shape. Fitting the local average response may provide a more accurate model of the relative activity of voxels across a sphere as compared to fitting a fixed response function across the whole brain.

Initial voxel selection

Data analysis was performed on the entire scanned brain volume with subregions defined by the sequential searchlight. To distinguish the brain, in its entirety, from the surrounding skull and empty scanner space, a voxel mask was applied based on functional data using standard AFNI procedures. Voxels outside the full-brain mask were set to 0 at all time points; these 0 values were incorporated into searchlight analyses when performed close to the exterior of the brain. Because the inclusion of these null values was consistent across all stimuli, it did not affect the patterns of the dissimilarity matrices.

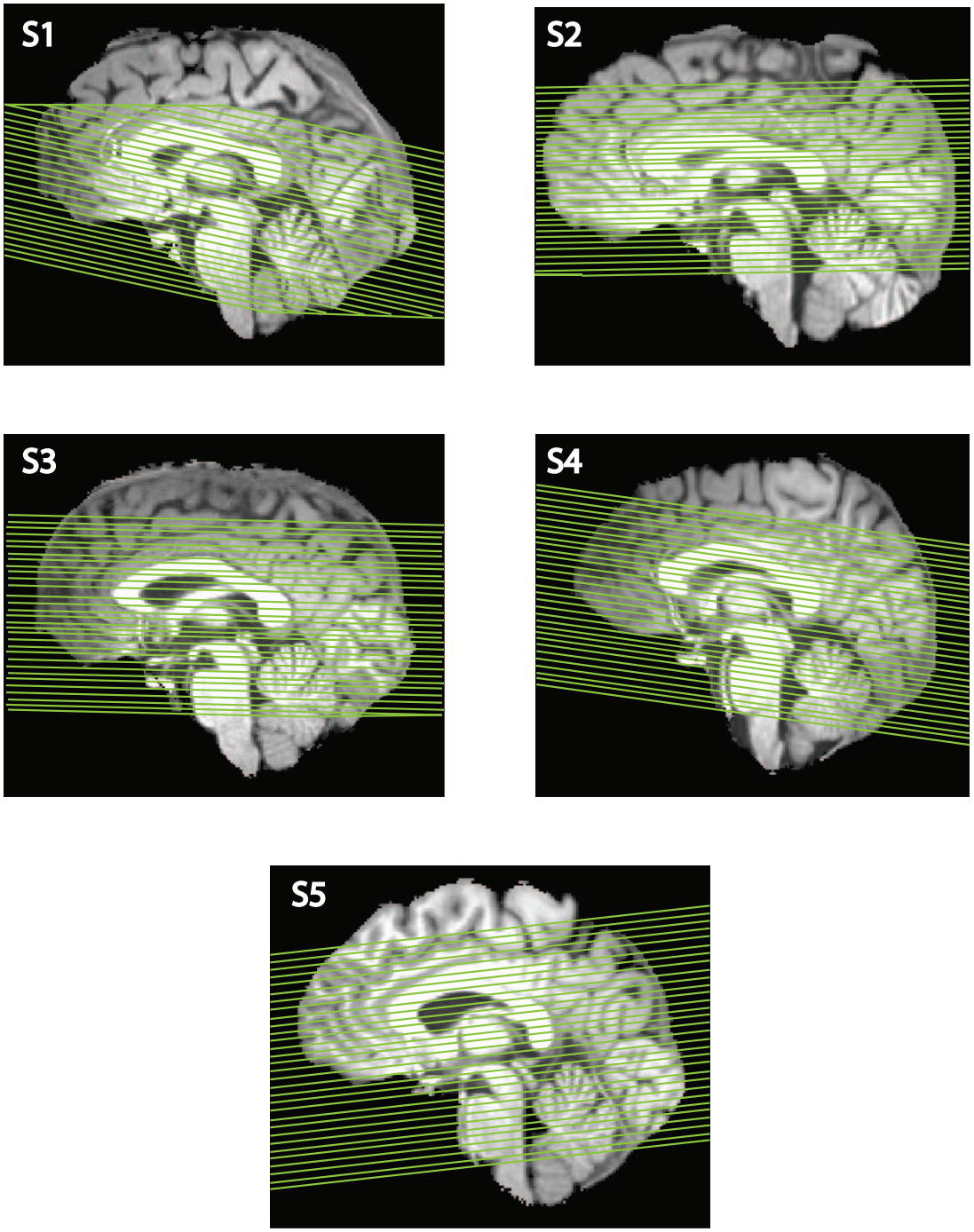

Figure 2. Slice coverage for all subjects.

Representational dissimilarity measures

As discussed earlier, we employed *representational dissimilarity* as a means for relating the neural representation of objects to the representation of the same objects within a variety of computer vision models. A *representational dissimilarity matrix* (RDM) D^m was computed for each encoding model m such that

$$D^m_{i,j} = d^m(s^i, s^j) \quad (1)$$

meaning the matrix element in the i^{th} row and j^{th} column contains the distance, or *dissimilarity*, between the i^{th} and j^{th} stimulus s^i and s^j in the model m. A given dissimilarity matrix captures which visual objects are clustered together by the corresponding representation. The searchlight procedure was then used to identify voxel clusters with D^ms similar to the RDMs of each computer vision model.

A 123-voxel searchlight sphere was defined centered on each voxel in the brain (Kriegeskorte et al., 2006)

with individual voxel responses to each stimulus computed as described in the Preprocessing section. For a given searchlight centered on voxel location (x, y, z), each RDM entry $D_{i,j}^{searchlight_{x,y,z}}$ was defined as one minus the Spearman correlation between the voxel responses for stimuli i and j (Kriegeskorte et al., 2008):

$$d^{searchlight_{x,y,z}}(s^i, s^j) = 1 - r(v(s^i), v(s^j)) \quad (2)$$

The 123-element vector $v(s^i)$ represents the voxel responses for stimulus i averaged across all six blocks to compute the RDM. This averaging enhances the stimulus-specific response over the independent time-varying noise, providing a more stable estimate of the searchlight response to each stimulus.

Five computational models of object representation were implemented for comparison with the neural representation of objects. Four of these methods were drawn from popular computer vision models with varied approaches to object representation, and the fifth was a standard computational model designed to account for neural responses relatively early in the primate visual cortex. Distinct distance metrics $d^m(\cdots)$ were derived from each method. These models, ordered from relatively more local to more global feature representations, are described next.

Gabor filter bank

The Gabor filter is a well-established model of cell and voxel-level selectivity in V1 (Daugman, 1985). Each filter identifies a specific local-oriented edge in the stimulus. A bank of filters spans edge angle, position, and size. The first four levels of the filter bank used in Kay et al. (2008) were implemented and used to represent each image. The real-valued responses for each filter were recorded in a vector. Euclidean distance was used to measure the difference between the vectors associated with each pair of images.

Geometric blur

Geometric blur uses local image properties at selected interest points. The relative locations of these interest points are included in the image encoding, thus incorporating more global geometric properties of each object. Feature vectors consist of pixel values regularly sampled in radiating circles around the interest point with the starting point for sampling being determined by local image statistics. Pixel values are blurred over space with increasing blur for higher-radius circles. This approach emphasizes precise details at each interest point and lower-resolution context from the surrounding region, similar to the decrease in spatial resolution away from the retina's focal point in early vision.

Interest points were selected randomly from edges found by a Canny edge detector (Canny, 1986). Features were extracted through an implementation of the algorithm described in Berg, Berg, and Malik (2005). For each pair of images, each interest point in one image (the image with fewer points) was matched with the point spatially closest in the second image. The dissimilarity for each pair of points was computed by taking the weighted sum of the negative correlation between the two feature vectors, the Euclidean distance between the points, and the change in circle orientation as defined in Berg et al. (2005). The final dissimilarity between images was found by summing the dissimilarities for all pairs of points. This incorporates both global geometric information and spatially sampled local image statistics.

Scale invariant feature transform

SIFT features (Lowe, 2004) have been widely used in computer vision systems for visual object recognition. This approach selects significant pixel patches of an image and captures associated visual properties that are invariant to a variety of common transformations, such as rotation, translation, image enlargement, and (potentially) changes in ambient lighting. More specifically, for a given image, *interest points* are identified, and a scaled, rotated frame is defined around each point. For each frame, a feature vector is computed to encode the local image properties, defined as coefficients of a pyramid of image gradients increasing in spatial scope. SIFT features were extracted from the 60 object stimuli using the VLFeat package for Matlab (Vedaldi, 2011) with default settings when not otherwise specified.

A *bag of features* approach was used to compare SIFT features for pairs of images (Nowak, Jurie, & Triggs, 2006). Conceptually, each of the SIFT feature vectors in each stimulus is categorized as one of 128 "words," and the words are consistently defined across all 60 images. Each image is then characterized by the frequency of each of the possible words. More specifically, k-means clustering is performed on the feature vectors from all interest points of all pictures, placing each vector into one of 128 groups. Assignment of multidimensional continuous-valued vectors to a small number of groups greatly reduces SIFT's representational complexity. A histogram is computed to find the frequency of each vector group in each image, and the histograms were normalized to sum to one. For each image pair, the Kullback-Leibler (KL) divergence was used to measure the difference between the resulting two normalized histograms.

Shock graphs

The shock graph provides a complete and unique representation of a given visual object's external shape

by constructing a modified form of Blum's medial axis (Kimia, Tannenbaum, & Zucker, 1995) based on the object's silhouette. The graph is a set of vertices, edges, and *shock* labels, $\mathcal{G} = (V, E, \lambda)$. Each vertex represents a point or interval along the medial axis, edges connect spatially neighboring points or intervals, and each label specifies the curvature of the portion of the silhouette associated with the corresponding vertices:

- $\lambda = 1$ when curvature is monotonic; object only widens or only narrows over an interval
- $\lambda = 2$ when curvature reaches a local minimum at a point; object narrows prior to the point in the axis and widens after the point
- $\lambda = 3$ when curvature remains constant over an interval; object silhouette ends in a semicircle or object is a circle
- $\lambda = 4$ when curvature achieves a local maximum at a point; object widens prior to the point in the axis and narrows after the point

Further details are provided by Siddiqi, Shokoufandeh, Dickinson, and Zucker (1999). The distance between graph pairs was computed using a graph-matching technique implemented by ShapeMatcher 5.2.1, which also was used to generate the graphs (Macrini, 2008).

Scene gist

Although scene gist (Oliva & Torralba, 2001) is specially designed for recognition of scenes rather than objects, we included this model partly as a control for our assumptions about object representation and partly to explore whether global image encoding methods are applicable to biological object perception. In the scene gist model, each picture is represented as a weighted sum of bases, found through PCA such that a small number of bases can be added together to reconstruct natural scene images with low error. The weights used in summing the bases to reconstruct an image serve as the features.

A scene gist feature vector for each image was computed using Matlab code implemented by Torralba (2006) and normalized to sum to one. The distance between each image pair was calculated as the KL divergence between the corresponding normalized feature vectors.

Two additional models and associated distance metrics were implemented as controls for the five computational models. The first control computed the Euclidean distance between grayscale *pixel intensities* of image pairs, comparing each image pair on the simplest visual level, absent of assumptions about perceptually meaningful visual properties. The second control computed the correlation between binary *semantic properties* of object pairs for 218 semantic features employed by Sudre et al. (2012), comparing objects largely based on nonvisual properties, such as animacy and edibility.

After defining the distance metrics and calculating the RDM entries for each of the five models, the resultant matrix for each model was compared to the matrix for each searchlight volume by converting the lower triangle of each 60×60 matrix into a 1770×1 vector and measuring correlations. When a model represents a set of image pairs as similar and a voxel sphere encodes the same pairs of images as similar, we may consider the voxels to be selective for the visual properties captured in the model. By comparing each computational representation with searchlights swept across the whole brain, we can identify which cortical regions, if any, have responses well described by each method's object/image representational approach.

Statistical significance values were computed at each searchlight location through permutation tests. The elements of the vectorized computer vision RDMs were permuted 500 times; the mean and variance of correlations for each searchlight position with each permuted RDM were computed to derive z values for the true correlation measures. The z values were converted into one-tailed p values, using $p = (1 - erf[z])$, where erf is the cumulative density function of the Gaussian distribution $\mathcal{N}(0,1)$. A threshold was chosen for each subject and each computational model such that the false detection rate was $q \leq .001$, typically $p \ll .001$, following the method of Genovese, Lazar, and Nichols (2002), and the regions above threshold were visualized over the subjects' anatomical surfaces. Surface maps were constructed using *FreeSurfer* (2012) and SUMA (Saad, 2006).

Results

Our study was designed to illuminate how the human visual system encodes object information along the ventral pathway and, in particular, explicate the nature of intermediate neural object representations. To that end, we employed five computational models that make specific, and different, assumptions about the algorithms for recognizing content in visual images (see the Representational Dissimilarity Measures section). To the extent that there is a gap in our knowledge with respect to the nature of intermediate features in human vision, we adopted these models as proxy theories, which each provide differing constraints on possible representations. Individual models were compared to our fMRI data by measuring the distance or *representational dissimilarity* between each pair of object stimuli for both the particular computational model and the neural encoding. A searchlight method was used to

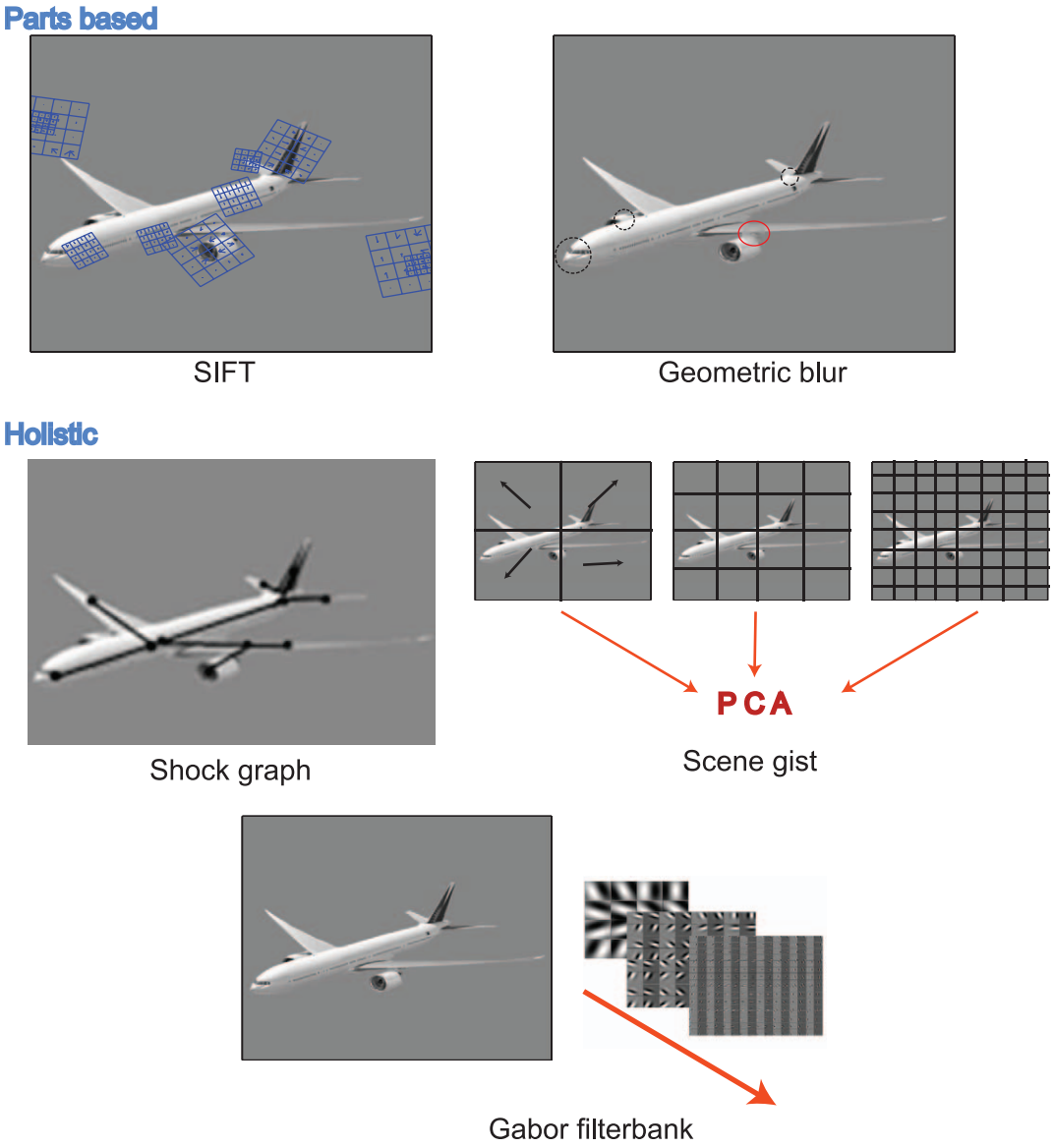

Figure 3. Five computational models under analysis. Features from SIFT are edge statistics measured at selected local scaled and rotated frames; features from geometric blur are blurred pixel values circularly sampled at local interest points; features from shock graph are computed based on shape of object silhouette; features from scene gist are edge statistics sampled at multiple scales across the full image and compared against standard scene statistics; features from Gabor filter bank are projection of full image on Gabor wavelet pyramid.

identify brain regions in which the set of interstimulus distances, that is, the grouping of the stimuli, was similar to the grouping of the same stimuli produced by a given computational representation. Of note, in comparison to the limited functional regions identified by the LOC localization technique discussed in the Experimental Design section, we searched almost the entire brain to allow for the existence of brain regions selective for complex visual features beyond those regions often associated with object representation.

Given that all five of our included models rely on the same visual input as our fMRI experiment, it is not surprising, but still gratifying, that we observe significant correlations between our neural data and all five models. Figure 5 depicts those brain areas with significant correlations ($q < .001$) between the distance matrices derived from each model and the neural responses within each area. Importantly, although we scanned across almost the entire brain, these correlated brain areas are focused in anatomical locations associated with low-, mid-, and high-level vision in both dorsal and ventral visual cortices with limited spread to the prefrontal cortex (PFC). Overall, the SIFT model most consistently matched the obtained stimulus representations in midlevel visual areas, and the Gabor filter bank model most consistently matched the obtained stimulus

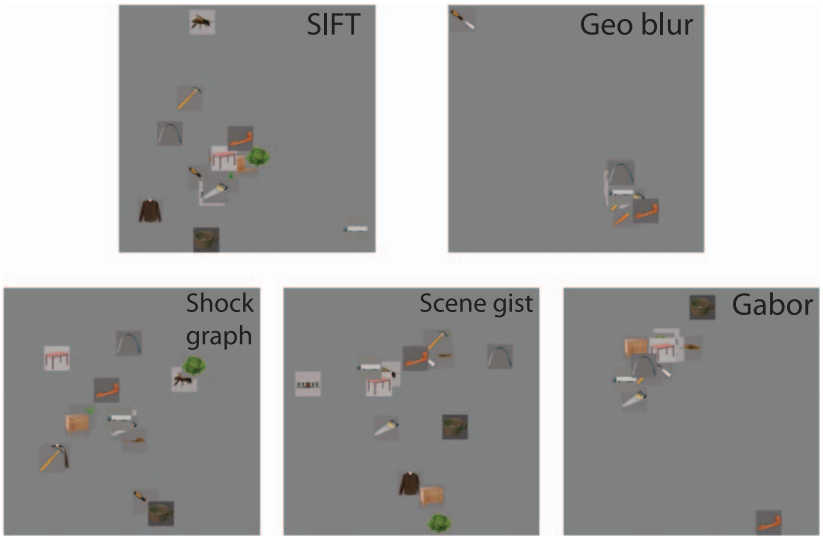

Figure 4. Multidimensional scaling visualization of the relative clustering of 15 of the stimulus pictures based on each computational model under analysis.

representations in low-level visual areas. The neuroanatomical locations for matches to the three other models were less consistent across subjects.

If we consider the underlying characteristics of each model, these results appear reasonable. First, the Gabor filter bank model encodes local oriented edges and has been used successfully to model receptive fields in the early visual cortex (Kay et al., 2008). Thus, the distance matrix correlations resulting from the Gabor filter bank model serve as a baseline to ensure that our overall approach is viable. As such, we expected a significant match between the activity observed in human V1 and this model. Moreover, including the Gabor filter bank model allows us to contrast these baseline correlations expected to be associated with earlier visual processing with any observed correlations arising in mid- and high-level visual areas. As illustrated in Figure 5 in orange, S2, S3, and S5 all show a positive correlation between the RDMs from the Gabor filter bank model and neural activity in the left occipital pole, and all five subjects show a positive correlation in the right occipital pole. Somewhat surprisingly, the Gabor filter bank model also elicits significant positive correlations in midlevel visual areas, including the left fusiform cortex (lFus) in all five subjects and the right fusiform cortex (rFus) in subjects S2, S3, S4, and S5; subjects S2, S3, and S5 also exhibit positive correlations in the left LO. We also see some correlation in anatomical regions often associated with higher-level visual processing, for example, extending more anteriorly in the ventral temporal cortex for S1, S4, and S5. Finally, the Gabor filter bank model is correlated with activity beyond the ventral stream, including the inferior parietal (IP) region in the left hemisphere of S2, S3, and S4, and in the right hemisphere of S2; somewhat smaller correlations were also observed in the left PFC of S2 and right PFC of S3 and S5. Least intuitive may be the small-area, weak-correlation matches in the left precentral sulcus of S3 and S5. Figure 6 emphasizes the most consistent match regions across subjects are in the bilateral occipital poles and early ventral stream.

In contrast with the Gabor filter bank model, the SIFT model encodes local visual statistics selected across points of interest in an image. The more restricted results observed for the SIFT model are consistent with this difference in representation. Positive correlations between the SIFT model and regions of neural activity are evident in subjects S2, S3, S4, and S5 as illustrated in Figure 5 in blue. With respect to the SIFT model, our major finding is that these four subjects all show positive correlations in the bilateral fusiform cortex. Subject S5 also shows a positive correlation in the bilateral LO. In the dorsal stream, there is strong positive correlation for S2 in the left IP. We also observed a positive correlation in the left PFC for S5 and right PFC for S2 and S5. Figure 6 illustrates the overlap of positively correlated regions across subjects in the bilateral fusiform cortex and in the posterior right ventral stream.

The geometric blur model, much like SIFT, encodes local visual properties from selected points in each image but also encodes more global information about object geometry. As illustrated in cyan in Figure 5, all five subjects showed positive correlations with neural activity in midlevel visual areas; the breakdown by subjects is illustrated in Figure 6. Subjects S1 and S5 exhibited positive correlations spanning the bilateral fusiform cortex and posterior IT (pIT) with S5 exhibiting a more continuous region. More anteriorly

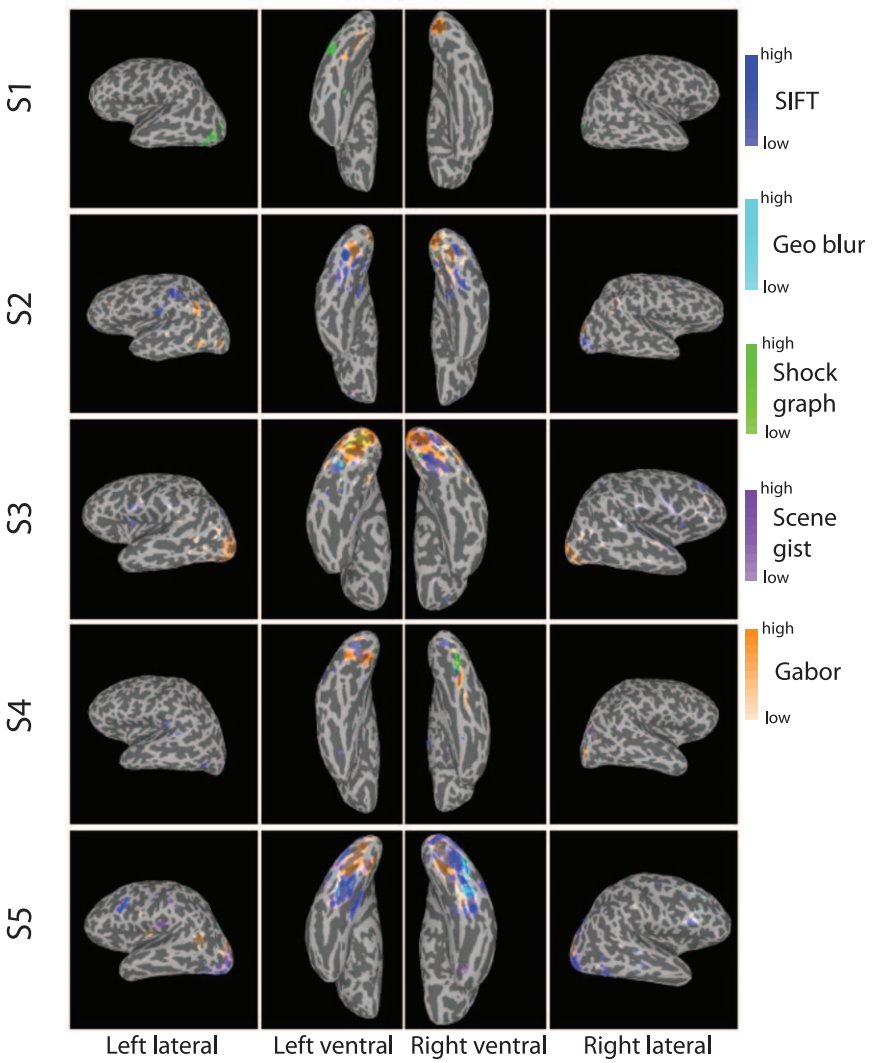

Figure 5. Cortical regions with a dissimilarity structure significantly correlated, $q < .001$, with the dissimilarity structures of the five different models of visual feature coding. Colors are associated as follows: blue for *SIFT*, cyan for *geometric blur*, green for *shock graph*, purple for *scene gist*, and orange for *Gabor filter bank*. Color intensity proportional to correlation. Regions matching multiple models show the corresponding colors overlayed. Note first that although we illustrate these results on surface maps, the actual searchlights were run on brain volumes and, second, that color overlap sometimes forms misleading shades, for example, purple as the combination of blue and orange. Compare with Figure 6 in cases of uncertainty.

in the right IT, we observed a spatially smaller positive correlation for S1 and S4. The right occipital pole also had small spatial regions showing positive correlations for S1, S2, S3, and S5 in addition to regions near the left occipital pole for S1 and S5. Within the ventral visual cortex, S5 also shows a positive correlation in bilateral LO. In the dorsal stream, there are small positively correlated areas in the parieto-occipital sulcus (POS) for S2. Finally, we observed a positive correlation in PFC for S5.

The shock graph model uniquely represents the silhouette shape of a given visual object, ignoring small-scale internal details critical to more local models, such as SIFT and geometric blur. Positive correlations between neural activity and the shock graph model are illustrated in green in Figure 5. These positive correlations are apparent for subjects S1, S3, S4, and S5. S1 exhibits positive correlations in the bilateral LO and bilateral occipital poles. There are positive correlations for S3, S4, and S5 in rFus as illustrated in Figure 6.

The scene gist model encodes global image properties most commonly found in natural scenes, focusing on the two-dimensional spectrum across a given image. Positive correlations for the scene gist model are shown in purple in Figure 5 with the most robust results being observed in S5 although, as illustrated in Figure 6, there are also positive correlations in S1, S3, and S4. More specifically, S1 and S5 exhibit positive correlations in lFus. S5 also shows positive correlations in the

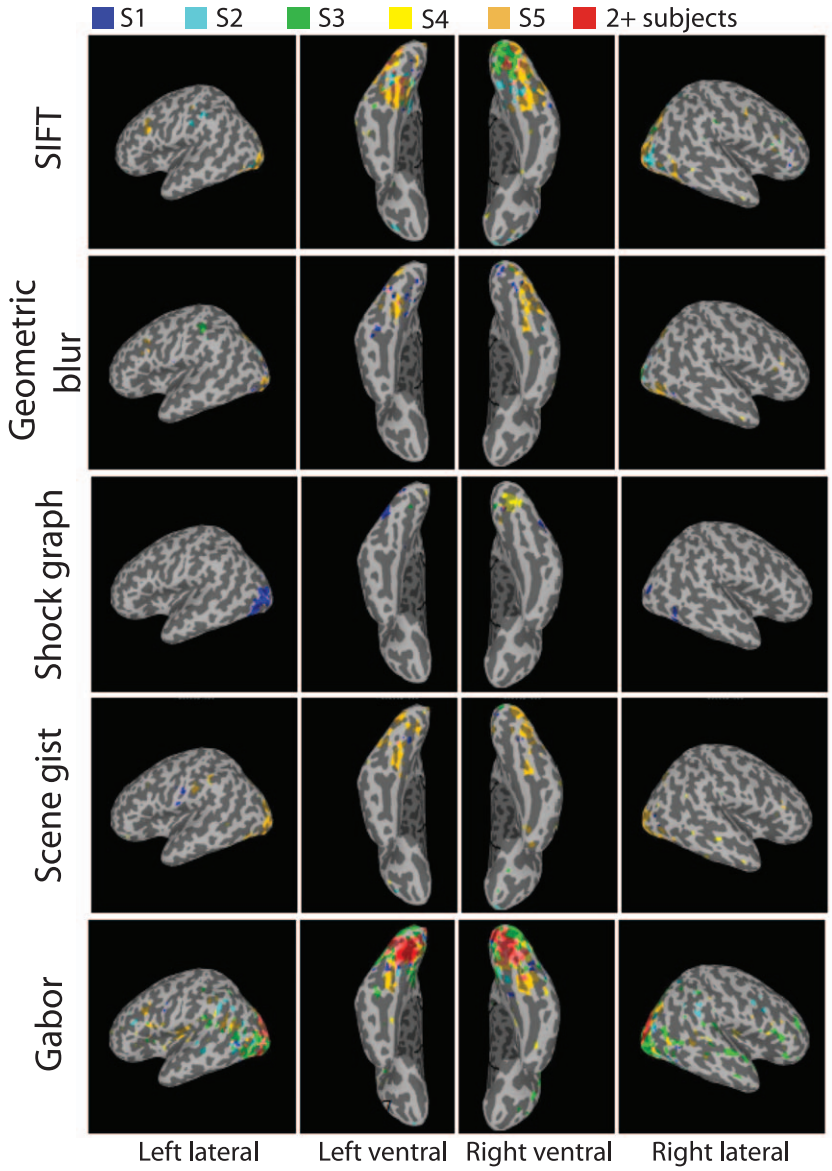

Figure 6. Cortical regions on Talairach brain with dissimilarity structure significantly correlated, $q < .001$, with the structures of computer visual models. Colors are associated with subjects as follows: blue for S1, cyan for S2, green for S3, yellow for S4, and orange for S5. Red denotes overlap between two or more subjects with darker shades of red corresponding to increasing numbers of subjects overlapping with one another.

rFus, bilateral LO, and bilateral pIT. S3 and S5 show positive correlations in the right occipital pole with S5 also showing a positive correlation in the left temporal pole. Less robust effects are seen for S4 and S5 in a more anterior region of the right IT; and S1 and S5 show positive correlations near the left IP.

Taking a somewhat broader perspective, comparisons among these results indicate that some brain regions appear to consistently correlate with several of the computational models we considered. First, the geometric blur and SIFT models, both encoding local statistics of images, have overlapping regions on the ventral surfaces of S3 and S5 and in the PFC of S5. Within the ventral surface, these regions tend to be in the pIT. The greatest degree of overlap can be seen between SIFT and the Gabor filter bank model across subjects S2, S3, S4, and S5, largely along the posterior ventral surface. To some extent, this may be expected simply by chance as these two methods produce the largest sets of model-cortical match regions. It also is worth noting SIFT is based on nonlinear operations on selected Gabor filter responses, potentially tying the two methods together.

Examination of searchlight correlation values pooled across broad, independently defined anatomical regions provides additional perspective on the descriptive power of each computational model with respect to activity in the visual cortex (Figure 8). Correlations for

the Gabor filter bank model are significantly higher than correlations for all alternative models across subjects and across a variety of primarily posterior regions, including the right occipital cortex and the left lingual cortex. At the same time, correlations are also present in anterior regions, such as rFus. Correlations for the SIFT model are significantly higher than correlations for all alternative models in the left and right occipital cortices for S2 and S5 and in the right middle occipital cortex for S5 as well as more anteriorly in lFus for S4 and S5, in rFus for S5, and in rIT for S4. Correlation comparisons for the other three models are rarely significant and scattered across the visual cortex.

The two control models, encoding particularly simple information in one case (pixel intensities) or particularly abstract in the other case (semantic properties), were considered as contrasts to the five computational models of visual perception. The locations of the two control models' few significant correlations with neural activity reinforce the use of the searchlight RDM method in the study of cortical visual representations.

The pixel-intensity model reflects the presence of bright or dark pixels at each location in an image, and the Gabor filter bank model of V1 discussed above reflects the presence of contrast edges between groups of pixels across narrow regions of an image. The algorithmic similarity between these models and the conceptual simplicity of the pixel-intensity comparison lead us to anticipate most regions showing significant positive correlations with the pixel-intensity model will appear early in the ventral pathway. This is observed in part as illustrated in Figure 9. S1 and S4 show relatively large regions of positive correlations bilaterally at or near the occipital pole, and S2 shows positive correlations near the right occipital pole. The pixel-intensity model also produces significant positive correlations in the right LO for S1, left anterior IT (aIT) for S2, and right aIT for S4, spanning smaller cortical regions. Less intuitively, there are small positively correlated areas in the left IP for S2 and S5, right IP for S3, bilateral POS for S5, and bilateral orbital frontal cortex for S5.

The semantic model encodes generally nonvisual information about each object, leading us to anticipate most regions showing significant positive correlation will appear at high levels of the ventral pathway or outside the ventral pathway entirely. As expected, as illustrated in Figure 9, the few regions of positive correlation appear more anteriorly, in the left LO and left Fus for S1 and in the bilateral Fus for S2. The model does not significantly match neural activity in any region for S3, S4, and S5. The lack of frequent positive correlations is not surprising as the semantic model is designed to compare 251 high-level properties simultaneously—unlikely to be localized to a single brain region.

Another way of examining this data involves focusing on a specific functional region—in this case, the area of the ventral stream most often associated with generic, high-level object processing, the LOC (Kourtzi & Kanwisher, 2000; Grill-Spector et al., 2001). Overlap between cortical regions differentially selective for objects, identified using the LOC "localizer" described above, and searchlight volumes found to be positively correlated with one or more of the five computational models are illustrated in Figure 10. These overlap regions were spatially small as compared to the overall volumes identified by the searchlight process and varied in anatomical location depending on the particular computational model and the subject. For example, within the LOC, the anatomically based left LO overlapped with a volume identified as correlated with the Gabor filter bank model in S3, and the lFus showed overlap with volumes associated with the Gabor filter bank model in S4. Further overlap within the LOC was observed for Gabor filter bank volumes located in the right pIT for S4, in a more anterior region of the left IT for S1, and in the left extrastriate cortex for S3 and S5. With respect to correlated searchlight volumes arising from the SIFT and geometric blur models, within LOC we observed overlap in the right LO, pIT, and more anterior IT for S5. Finally, the geometric blur model overlapped with LOC responses in the anterior IT for S1.

To provide perspective on the similarities among the five studied computational models, we compared their respective stimulus distance matrices in Figure 11. We compute correlations for distance matrices including 59 of the 60 rows and columns and observe the average and standard deviation for each model comparison. We observe that the correlation between models' stimulus grouping structures generally fails to act as a predictor of overlapping regions seen in Figure 5 with the potential exception of the link between SIFT and geometric blur. Figure 11 also illustrates that the models have notably low pairwise correlations, that is, representations, of the 60 stimuli. Supporting this observation, for the most part, there are few overlapping regions across models in any of the five subjects. The low correlation between SIFT and Gabor filter bank, despite the overlap of their corresponding match regions in neuroimaging data, may reflect the removal of image location-specific information encoded in the Gabor filter bank but discarded in the bag-of-features evaluation of SIFT descriptors (Nowak et al., 2006).

A distribution of the model-neural activity positive correlation values, akin to a Gamma distribution, is above the false discovery rate threshold for each subject and for each model. The nature of these distributions is illustrated in Figure 12. Note that while the average

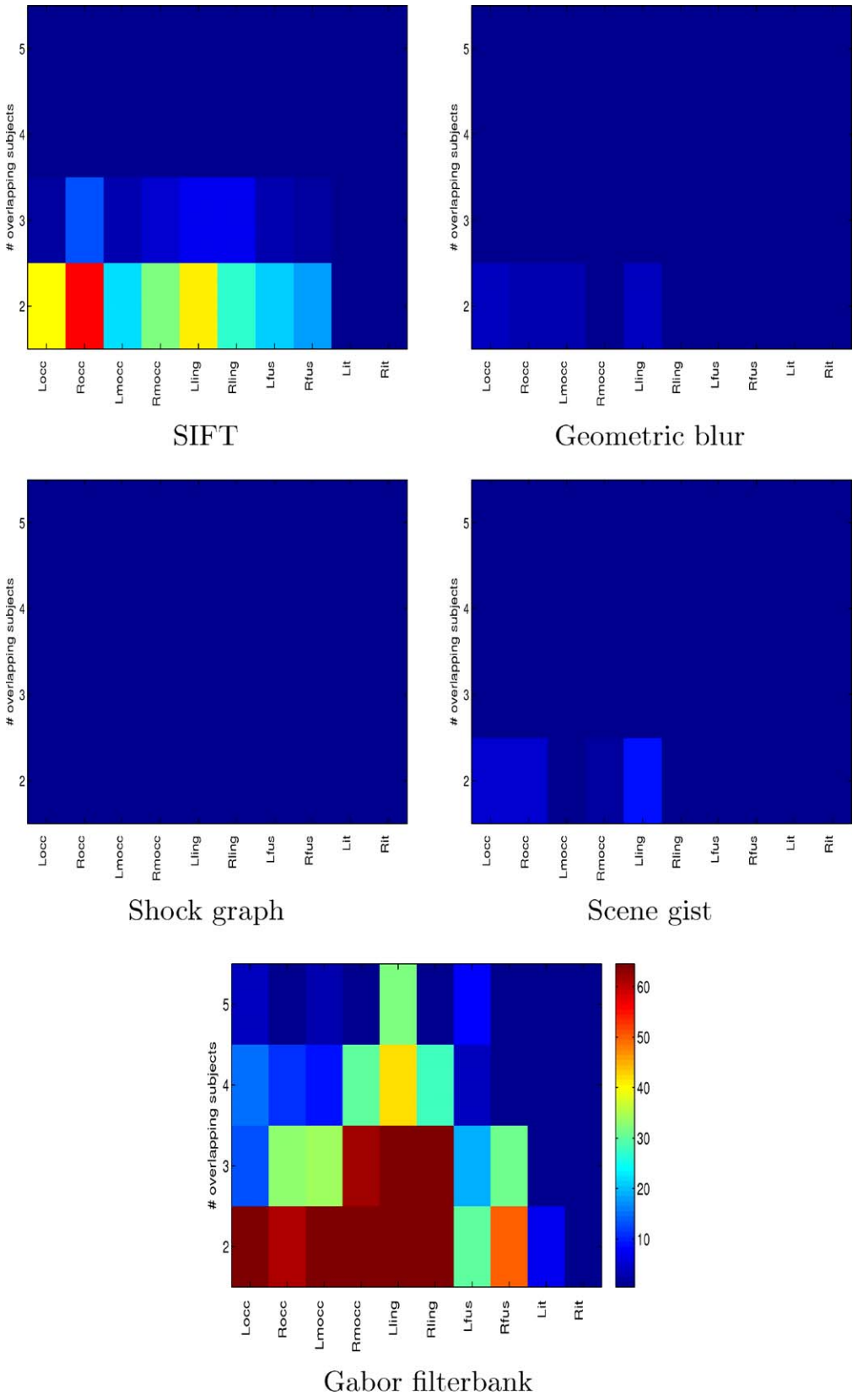

Figure 7. Number of voxel searchlight center in Talairach brain with more than one subject showing significant, $q < .001$, neural-model correlations. Results divided among anatomical regions of the ventral visual pathway. Number of searchlight centers represented according to color bar with dark blue for zero centers and dark red for 64 or more centers.

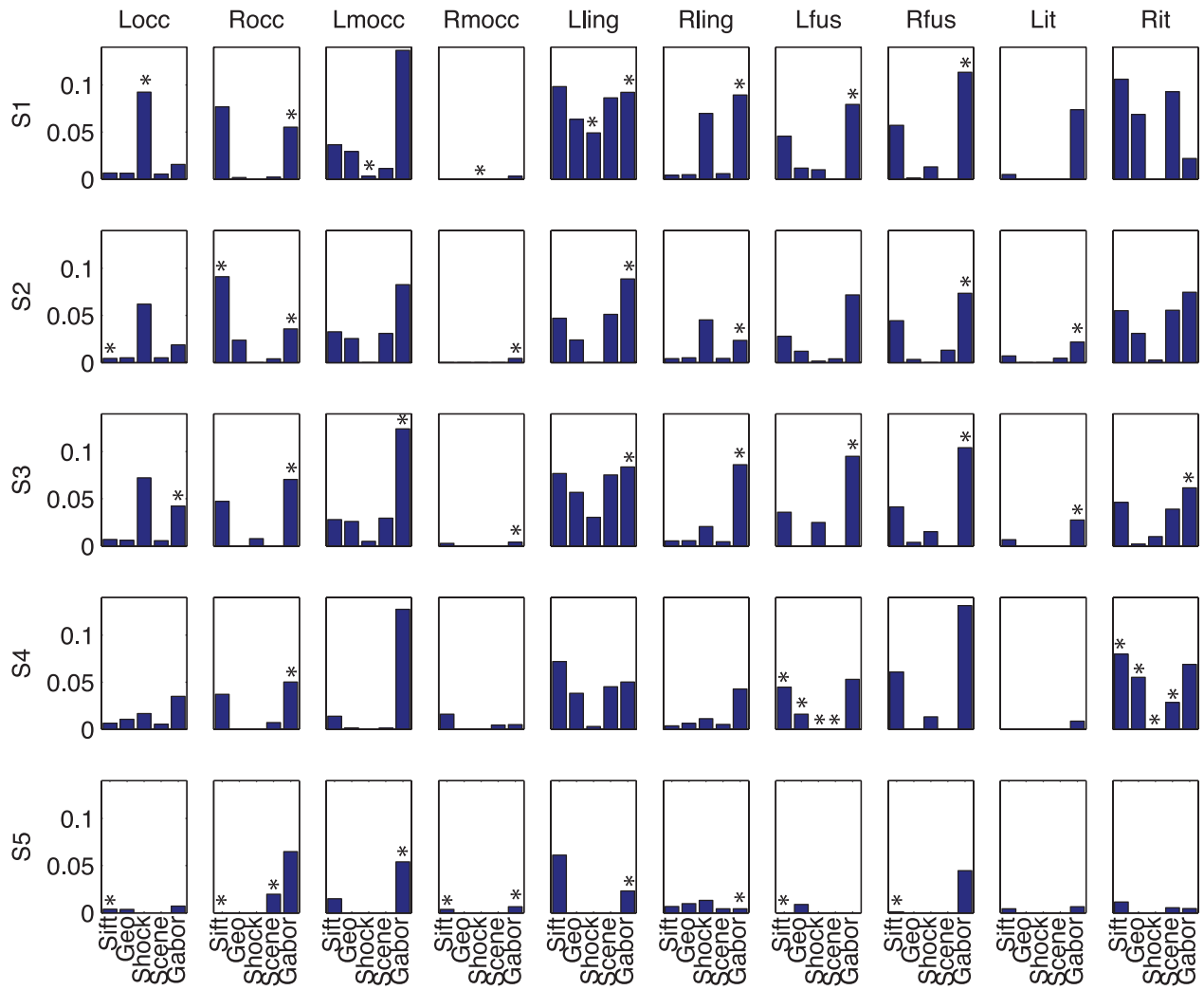

Figure 8. Average positive correlation value measured for searchlights in anatomical regions of the ventral visual pathway, based on Talairach projection. Star (*) indicates correlations in anatomical region significantly greater ($p' \leq 0.01$, Bonferroni-corrected for $5 \times 10 \times 5 = 250$ multiple comparisons) for the selected model than for the other four models.

significant correlations for each model are roughly the same, $r = 0.15$, the highest values provide a sense of ranking among computational representations in their abilities to account for neural responses. Most intuitively, the Gabor filter bank model, assumed to account for aspects of processing within the primary visual cortex, shows the strongest matches with an average top correlation of roughly $r = 0.33$; analysis of individual subject correlations reveals the same pattern. SIFT exhibits the second highest set of correlations with an average top correlation of roughly $r = 0.23$. The distribution of maximum correlations follows the same trend as the total area across all of the positively correlated regions for each model across all subjects; this is shown in Figures 5 and 6. Figure 12 also illustrates that there are significant positive correlations between every subject and every model. Certain matches are omitted from the discussion above because of their low correlations and their small surface spans, making them difficult to interpret.

The significant positive correlation values between model and neural RDMs are rather low—almost entirely below 0.3. To assess the extent to which noise in fMRI measurements adversely affected correlation magnitude, we computed and studied split half correlations. At each searchlight location, we recomputed the RDM based on voxel responses for the first three displays of each stimulus and, separately, for the second three displays of each stimulus. We then computed the correlation of the two RDMs at each location, shown in Figure 14. Split half correlations were markedly below 1.0. While they could rise as high as 0.66 for some subjects, most correlations were below 0.3 even in the ventral pathway presumed to respond most strongly and most consistently to visual object stimuli across displays. Thus, correlation values of 0.1

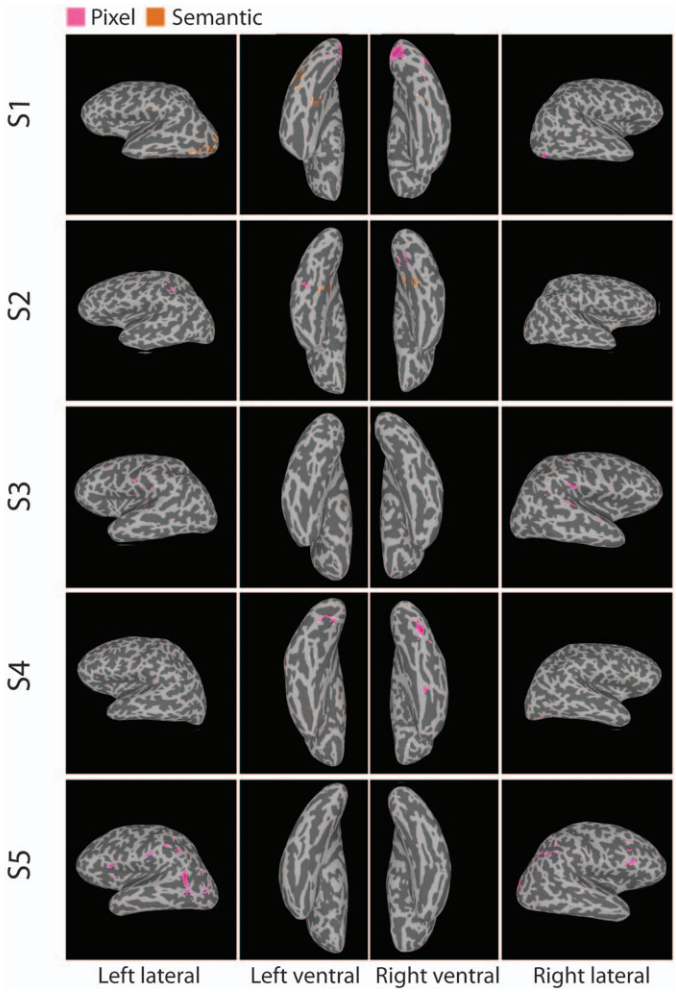

Figure 9. Cortical regions with a dissimilarity structure significantly correlated, $q < .001$, with the dissimilarity structures of five control models of object coding. Colors are associated as follows: pink for *pixel intensities* and brown for *semantic properties*.

and 0.2 may correspond to stronger matches in the absence of measurement noise.

The three-voxel searchlight radius used to compute each RDM was motived by parameter exploration reported in past literature to maximize neural encoding information (Kriegeskorte et al., 2006) and was further found using the present neuroimaging data to maximize RDM correlations. Figure 15 shows the correlation of voxel spheres centered at selected locations are at or near maximum value across subjects and models when using a searchlight radius of three voxels. In some conditions, increasing or decreasing the radius by two or more voxels could significantly alter the measured correlation value. Figure 16 suggests that additional image stimuli would decrease the variance of our measured correlations and hence increase statistical power. While the mean correlation is less dependent on the number of images than the number of voxels, from Figure 15, this is an expected property of dissimilarity matrices. Adding and removing images from a dissimilarity matrix corresponds to adding and removing rows and columns while leaving the remaining entries in the matrix intact. In contrast, adding and removing voxels, as in Figure 15, changes all of the values of the dissimilarity matrix. From the data we have collected, it is unclear how many images would be required for the reduction in measured correlation variances to saturate.

Discussion

Computational models of vision as proxy theories of biological vision

Our goal in this project was to better elucidate the featural selectivity of the ventral neural substrates supporting visual object processing. In contrast to our understanding of early visual processing (e.g., V1) and the high-level organization of visual cortex (e.g., the LOC, FFA, PPA, etc.), intermediate representation along the ventral pathway is poorly understood. To the extent that few theories account for this stage of visual object processing, we adopted a collection of theories drawn from computer vision to serve as proxies in that each theory makes specific, and different, assumptions regarding object representation.

To apply these theories to the neural representation of objects, we analyzed the pattern of similarity relationships between objects within the same collection of 60 objects as represented within the brain using patterns of object-generated activity recorded by fMRI and within each computational model. We then applied a searchlight analysis to uncover correlations between patterns of neural activity within brain subregions—sampled across the brain—and patterns within each computational model. This approach provided many regions in which there was a reasonable correspondence between a given model and the observed neural activity. Importantly, almost all of these significant correlations occurred in brain areas associated with visual object processing, thereby providing a theoretical sanity check that our results are informative with respect to our question of interest. At one level, this general result should not be particularly surprising; all of our models relied on the same spatial input, images of objects that were used as stimuli in the neuroimaging component of our study. Ideally, correlations at input should be reflected, at least to some degree, in correlations in representation of that input. On the other hand, the tested models each captured somewhat different linear and nonlinear structures in their representation of objects (e.g., Berg et al., 2005; Chandrasekhar et al., 2010). For example, the interest

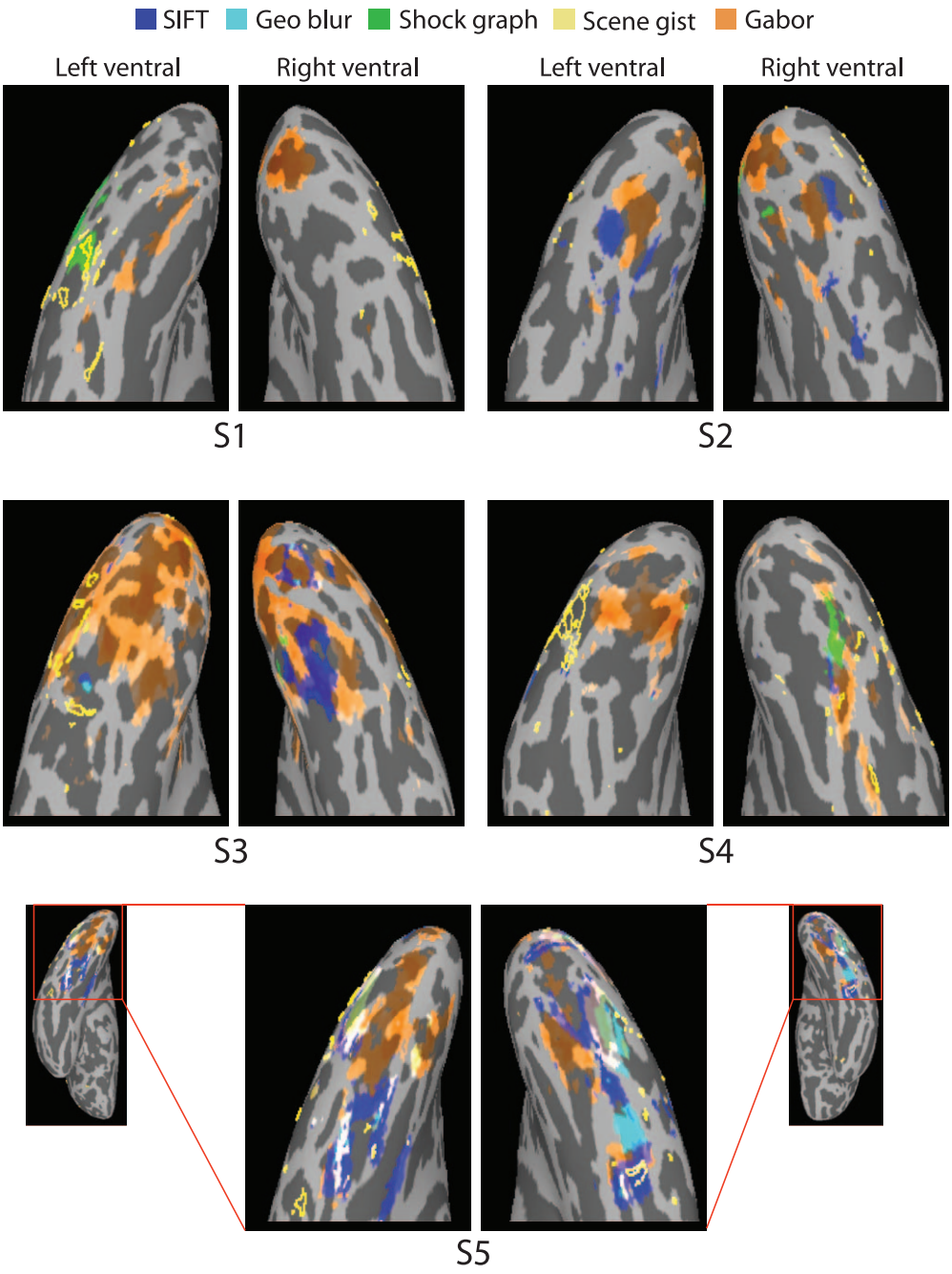

Figure 10. Cortical regions selected by LOC localizer and also found to have dissimilarity structure significantly correlated, $q < .001$, with the structures of computer vision models. Colors are associated as follows: blue for *SIFT*, cyan for *geometric blur*, green for *shock graph*, yellow for *scene gist*, orange for *Gabor filte rbank*. Yellow contours show LOC localized regions.

point frameworks used in the SIFT and geometric blur models provide a potential basis for parts-based perception—often assumed to be a critical element in the biological representation of objects (Schyns, Bonnar, & Gosselin, 2002; Zhang & Cottrell, 2005). In contrast, the shock graph approach provides a compact encoding of an object's silhouette and major axes, supporting a parametric description of holistic representation (Kimia et al., 1995). Scene gist is even more biased in representing global properties of an image, encoding the entire image structure of an object as well as its background (Oliva & Torralba, 2001). The pixel intensity and semantic property control models capture even more extreme low- or high-level object properties and are, accordingly, generally localized by our method to lower- and higher-level stages along the ventral pathway.

Beyond the basic finding that our highest model-neural response correlations are observed within the visual system, we gain further confidence regarding the

	SIFT	Geo blur	Shock graph	Scene gist	Gabor
SIFT	1	.39±.00	.04±.00	.48±.00	-.03±.00
Geo blur	.39±.00	1	.01±.00	.69±.00	-.09±.00
Shock graph	.04±.00	.01±.00	1	.05±.00	.04±.00
Scene gist	.48±.00	.69±.00	.05±.00	1	-.07±.00
Gabor	-.03±.00	-.09±.00	.04±.00	-.07±.00	1

Figure 11. Distance matrix Spearman correlations among the five models. Mean and standard deviation correlations computed using the leave-one-out method, leaving out 1 of the 60 stimuli for the distance matrices. Higher correlations in larger font and in darker red backgrounds.

informativeness of our method from the observation that the strongest correlations between the Gabor filter bank model and neural activity are located early in the visual pathway, near the occipital pole and extrastriate cortex. This finding is consistent with a wide variety of studies characterizing early neural visual receptive fields as coding for local-oriented edges (Hubel & Wiesel, 1968; Daugman, 1985; Kay et al., 2008). The extension of these significant correlation regions into the higher-level bilateral fusiform cortex and IP cortex has slightly less clear interpretations but may support the hypotheses of Cadieu et al. (2007) and Serre et al. (2007) that later stages of the ventral visual stream employ a hierarchy of sometimes nonlinear operations based on initial Gabor filter outputs. Beyond the operations specified in Serre et al. and Cadieu et al., SIFT represents a reframing of Gabor filter-like outputs for more complex recognition tasks, potentially accounting for the overlap in brain regions we observe between the correlations for the Gabor filter bank and SIFT models across subjects.

In summarizing the relative performance of the tested models, we find that both across and within subjects, the SIFT model appears to be the most promising of those tested for accounting for intermediate-level object representation in the human visual system. In particular, the SIFT model most strongly and consistently matched patterns of neural activity in rFus, an anatomical neighborhood associated with processing faces and other objects drawn from domains of expertise (Haxby et al., 2000; Tarr & Gauthier, 2000; Grill-Spector et al., 2004). To a lesser extent, we also observed correlations for the SIFT model within the left LO, a neuroanatomically defined brain region also associated with object perception (Grill-Spector et al., 2001). However, as shown in Figure 10, the SIFT model rarely correlates with brain regions lying within the functionally defined object-selective area referred to as LOC. Thus, it appears that the representation of objects in SIFT is similar to an intermediate encoding stage along the path to high-level object representation.

As a "proxy" model of intermediate feature representation, the preponderance of significant SIFT correlations in our results invites further reflection on its underlying algorithm. As discussed earlier, SIFT's interest point strategy is consistent with parts- or feature-based models of object perception. Notably, unlike geometric blur, our implementation of SIFT disregards the spatial locations of the local image regions it encodes, a characteristic that is consistent with the observation of invariance between intact images and their block-wise scrambled versions (Vogels, 1999). Similarly, SIFT incorporates aspects of the Gabor filter bank model, which does a reasonable job at capturing characteristics of early visual processing; as such, this component of SIFT enhances its nominal biological plausibility. Finally, our "bag of features" implementation of the SIFT model (Nowak et al., 2006) supports the learning of commonly occurring local edge patterns as "visual features." The use of such features allows the extraction of statistical patterns in the input similar to how vision scientists often characterize V1 receptive fields (Olshausen & Field, 1997). Each of these many algorithmic elements contribute to the stimulus representations embodied in our use of SIFT; further work is required to understand which of these elements contribute to the observed significant model-neural RDM correlations.

Our results also suggest that the shock graph model may be informative with respect to intermediate feature representation. Shock graphs describe objects in terms of their global shapes, capturing their axial structures and silhouettes. Thus, spatial information about the relative positions of shape features are preserved, but the local image statistics that may specify local features are not captured (e.g., texture). Our observation of correlations between ventral stream neural activity and the shock graph model supports the idea underlying shape-based encoding in intermediate-level neural representations (Tanaka, 2003; Yamane et al., 2008; Hung et al., 2012). To the extent that these correlations are confined to more posterior parts of the ventral stream, they are, however, somewhat inconsistent with the observation of Hung et al. of shape-based representations in the anterior IT in monkeys. At the

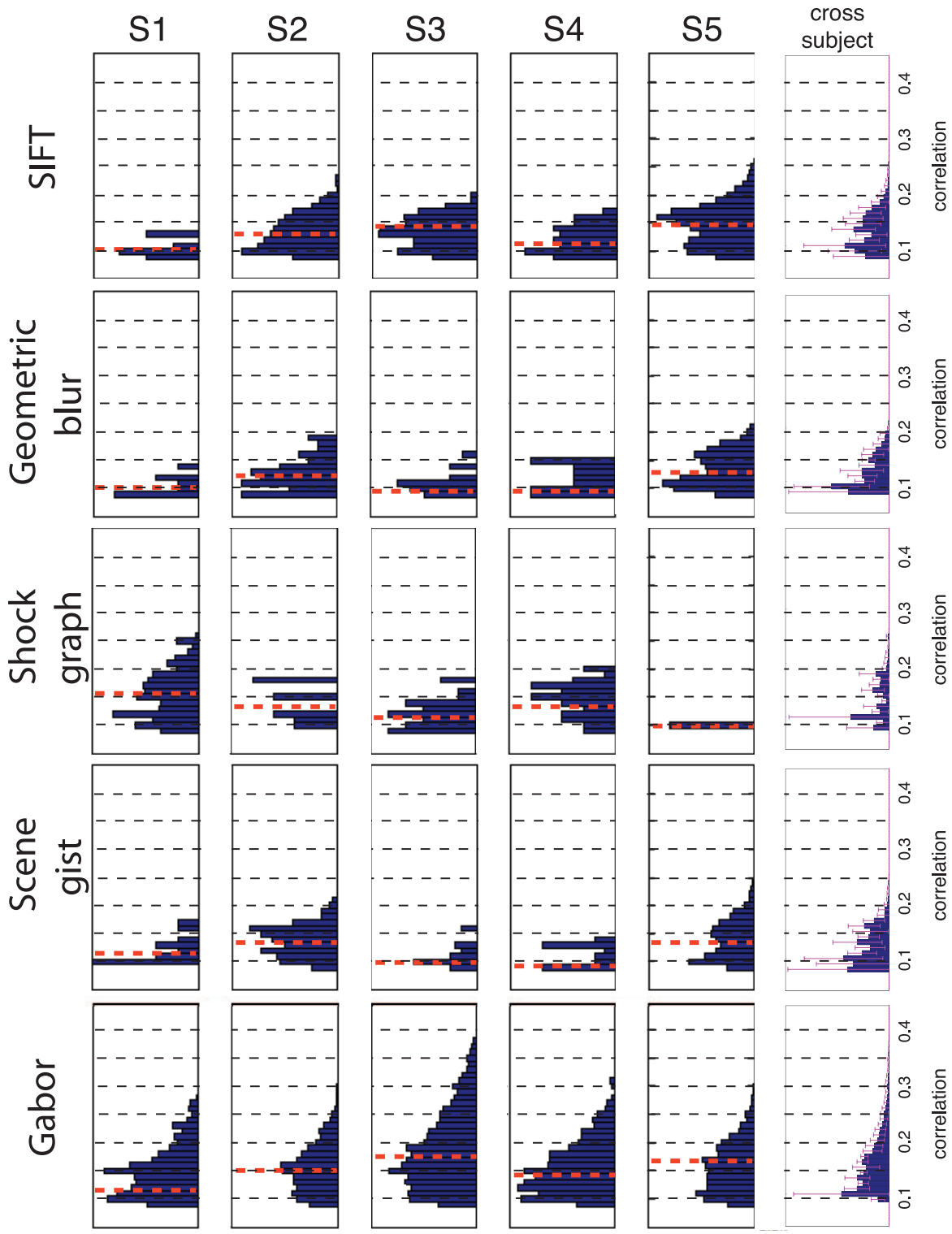

Figure 12. Histograms of significant correlations between model and searchlight RDMs. Median correlations for each model and each subject are shown as dashed red lines. The rightmost column shows the average histograms of the correlation values pooled across the five subjects, and each blue histogram bar is computed as a fraction of the total count; pink bars show the standard deviation for each correlation value bin.

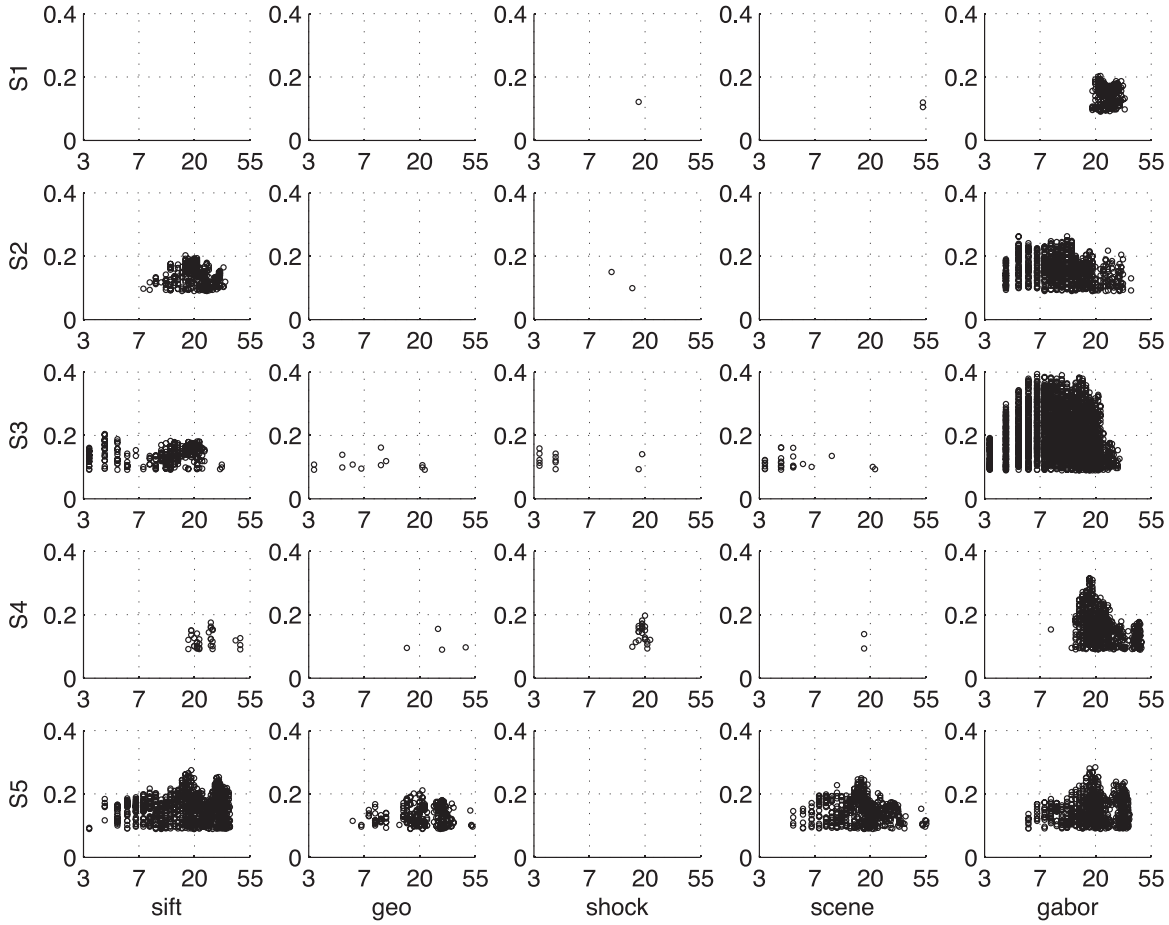

Figure 13. Scatter plot of significant correlation values (y axis) for each model by subject. Each plot depicts these correlations relative to the distance, in voxels, from the back of the brain, moving progressively in an anterior direction (x axis; log scale).

same time, this observation should not be generalized to other models of global encoding as we find that scene gist, encoding spatial frequencies across whole images, produces correlations in more anterior IT.

More generally, although our results are informative in some respects, it is doubtful that any established computational vision model accurately captures the neural representations instantiated in intermediate-level biological vision. Indeed, the best correlations between any model and the fMRI-derived cortical distance matrices (Figure 12) fall below 0.5. Nonetheless, the large majority of statistically significant ($q < .001$) model-fMRI correlations were found in visual brain areas with some differentiation within these areas for different methods. Furthermore, significant correlations generally were larger than pairwise correlations observed between model-derived distance matrices (Figure 11) and often approached the split half correlation values of voxel searchlight RDMs, indicating that some seemingly low values may be due to noise. Thus, while our conclusions are necessarily limited by the noise inherent in fMRI measurements, we suggest that our methods provide some sense of the visual properties for which given brain regions may be selective.

From a theoretical perspective, one potential concern with this interpretation is how we selected particular computational models for use in our study. In large part, our choices were based on each model's success or popularity in the extant computational vision literature and on each model's distinct encoding strategy with respect to intermediate feature representation—an intuition validated by the fact that the models have measurably different stimulus dissimilarity matrices (Figure 11). Of note, our present work does not include an analysis of the popular hiearchical model of biological vision known as "HMAX" (Riesenhuber & Poggio, 1999; Cadieu et al., 2007; Serre et al., 2007). HMAX employs a hierarchical series of summing and nonlinear pooling operations to model midlevel visual regions, such as V2 and V4. However, the HMAX model contains a variety of variables that must be fit either to the input stimulus set or to a set of experimental data (Serre et al., 2007). In an additional experiment not presented here, we found the actual data set collected in our study using the 60-image

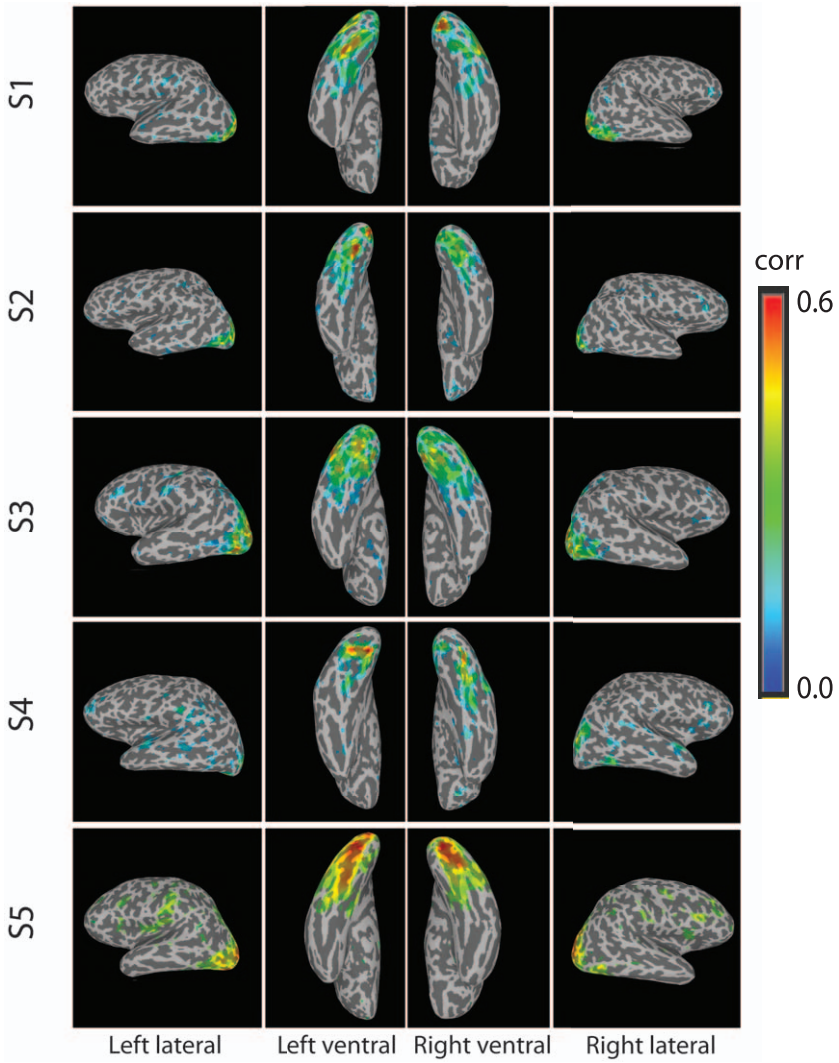

Figure 14. Split half correlation of searchlight RDMs based on the average of the first three recorded responses to each stimulus versus RDMs based on the average of the second three recorded responses to each stimulus. Correlations greater than 0.1 are displayed.

stimuli was insufficient for reliable fitting of HMAX (Seibert, Leeds, Pyles, & Tarr, 2012) even when limiting the model to layers S1 through C2 as in Cadieu et al. In contrast, the application of HMAX to the responses of individual neurons in monkeys (Cadieu et al., 2007) is more feasible as data for thousands of trials can be acquired. At the same time, it is worth noting that neurophysiological recordings of IT do not correspond to HMAX predictions for stimulus grouping structure (Kriegeskorte et al., 2008).

HMAX, considered up to the S2 layer, and the histogram of oriented gradients (HoG) model (Dalal & Triggs, 2005) both constitute prospective models of visual representation. Moreover, both are similar in form to the successful SIFT method. All three of these models rely on a nonlinear combination of local edges to characterize an image. HMAX computes local edges using Gabor filters in the "S1" layer (as in SIFT), records the maximum response for each filter orientation over a broader region within the given image in the "C1" layer, and computes weighted sums across the selected C1 edge responses in the S2 layer (Riesenhuber & Poggio, 1999). HoG computes local edges through computation of visual gradients of Gaussian-smoothed image patches and accumulates the gradients centered at each pixel through histograms normalized among broader patches (Dalal & Triggs, 2005). Conceptually, each model posits biologically plausible nonlinear pooling operations of local image properties that may relate to cortical object representations. Given the present results for SIFT, it seems likely HoG and the S2 layer of HMAX may also show significant correlations, perhaps in slightly differing brain regions. Of particular interest, weakening or strengthening of RDM correlations in response to changes in edge computations and pooling techniques may allow for a more clear

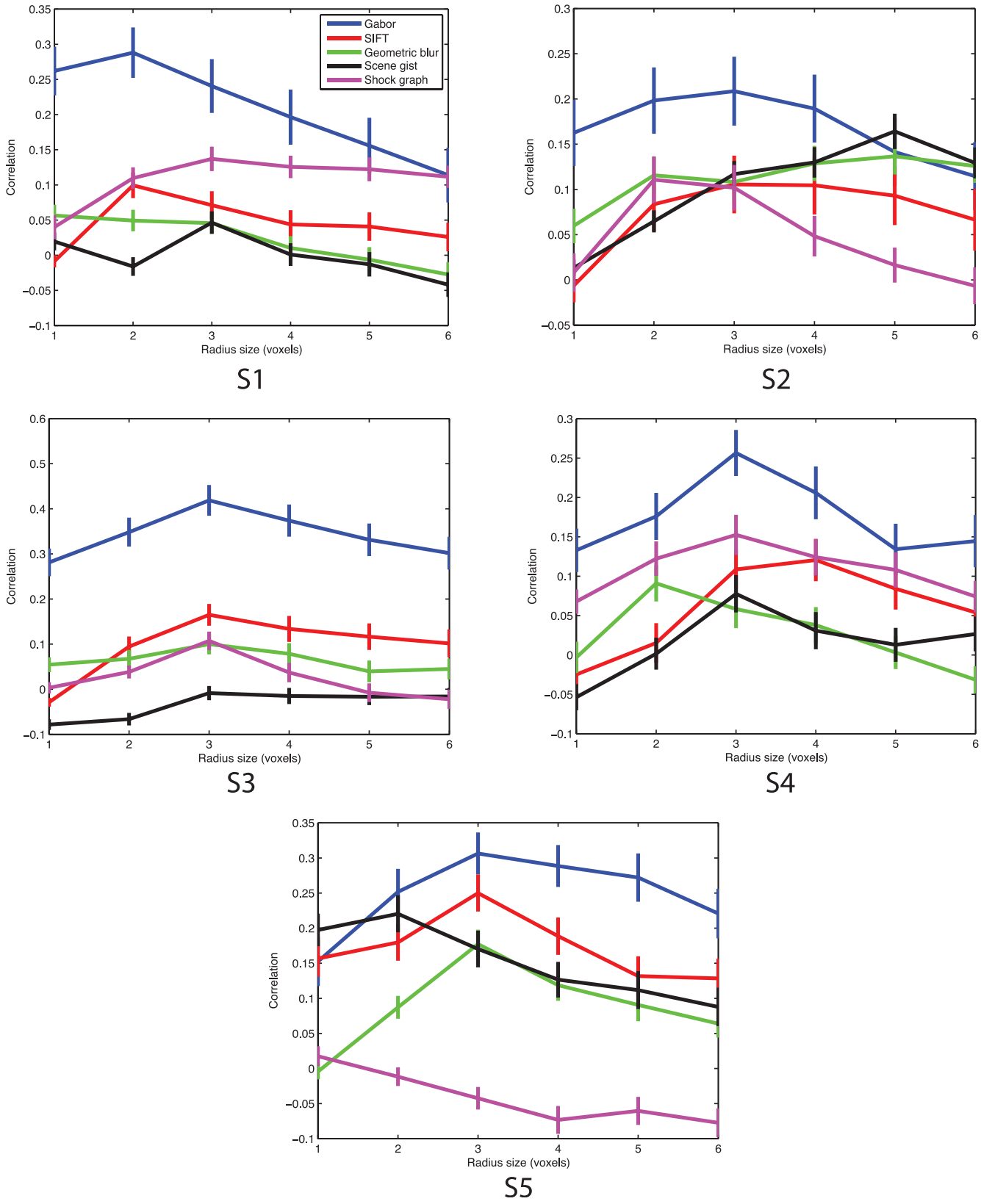

Figure 15. Correlations between RDMs for searchlights of varying voxel radii with model RDMs. The center for each searchlight was selected to be the voxel with the largest $r = 3$ searchlight correlation value for a given subject and model, provided there was sufficient room from the ends of the recording volume to also compute a $r = 7$ searchlight. Error bars were computed by correlating model and searchlight RDMs for 55 out of 60 images randomly selected 500 times. Colors are as follows: blue for *Gabor filter bank*, red for *SIFT*, green for *geometric blur*, black for *scene gist*, and purple for *shock graph*.

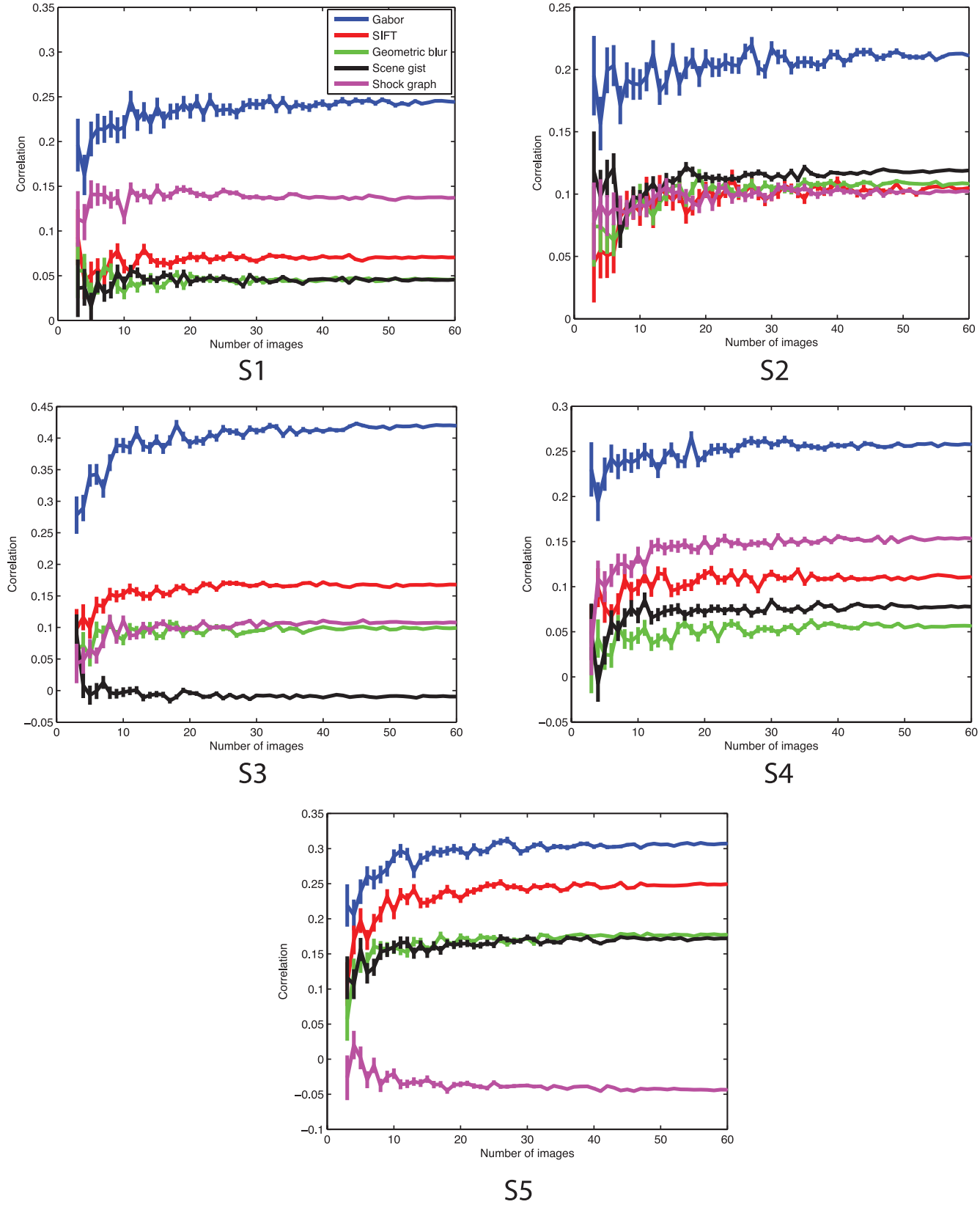

Figure 16. Correlations of each model and neural RDM using a variable number of stimuli. The location of each searchlight was selected to be the same as in Figure 15 for ease of comparison. Colors are as follows: blue for *Gabor filter bank,* red for *SIFT,* green for *geometric blur,* black for *scene gist,* and purple for *shock graph.*

understanding of the representational principles used in the ventral pathway.

From an empirical perspective, a second potential concern is the degree of variability in the spatial location, or even the existence, of large high-correlation brain regions for each model within individual subjects. In some cases, as in SIFT and Gabor filter bank, the changes in anatomical positions across subjects were relatively slight, consistent with variability of functional region locations, such as LOC or FFA (Kung, Peissig, & Tarr, 2007). More qualitative variability, for example, across lobes or hemispheres, may reflect meaningful differences in our subjects' cognitive and cortical approaches to object perception. For example, individuals may vary in the degree to which they attend to local versus global features or apply holistic mechanisms (Wong, Palmeri, Rogers, Gore, & Gauthier, 2009). Beyond the potential strategic variation in how individuals perceive objects, noise in the hemodynamic signals may increase the variability of correlated brain regions across subjects. However, this latter possibility fails to explain why all subjects exhibit significant and consistent correlations within the visual pathway for several of the models.

Conclusion

Our study aims to connect the cortical encoding in mid- and high-level visual areas of the human ventral stream and image representations as realized in several different computational models of object representation. Perhaps the most salient conclusion we can make is that the best biological/computational correspondence is observed for the SIFT model (Lowe, 2004). Although such results do not imply that SIFT-like computations are somehow realized in the human brain, they do suggest that the SIFT model captures some aspects of the visual stimulus that are likewise instantiated in human visual object processing. In particular, SIFT pools information about the presence of oriented edges distributed across image space, an approach generally understood to be used in cortical perception. Further study is required to assess which representational elements of SIFT are most biologically relevant.

As this study is one of the first attempts to directly connect extant computational models of object representation with the neural encoding of objects, there remains ample room to sharpen our observations and to further explore the space of possible biological vision representations. For example, maximum possible correlations between model and cortical representations of stimulus objects presently are rather low—generally below $r = 0.3$. Our results are limited by measurement noise and relatively weak hemodynamic signal measured to reflect cortical activity. To strengthen future studies, the passive viewing task used in the neuroimaging component of our study could be replaced by an active object identification task, which, conceivably, might yield stronger neural signals and more robust results. In contrast, broadening the set of stimuli viewed to enable a wider comparison of visual representations would offer little additional benefit in the absence of a dramatic increase in the number of stimuli and, likely, in the required scanning time.

Another improvement would be to consider other computational vision models, for example, HoG (Dalal & Triggs, 2005), the more biologically inspired model HMAX (Riesenhuber & Poggio, 1999), or the hierarchical model described in Jarrett, Kavukcuoglu, Ranzato, and LeCun (2009). SIFT's similarity to the HMAX S2 layer and to HoG—all three models rely on nonlinearities to pool local edge information—indicates further pursuit of these kinds of representations to describe high-level voxel encodings may prove a fruitful course for future research. Finally, a more sophisticated approach to developing informative model-brain correspondences may be to combine the dissimilarity matrices for any group of representational methods with weights optimally learned to match the representation within any given brain region (Tjan, 2001).

In sum, our study provides a foundation for further exploration of well-defined quantitative models using dissimilarity analyses and points the way to methods that may help shed further light on the visual features encoded in the human brain.

Keywords: neuroimaging, object recognition, computational modeling, intermediate feature representation

Acknowledgments

Research was funded by NIH EUREKA Award #1R01MH084195-01 and the Temporal Dynamic of Learning Center at UCSD (NSF Science of Learning Center SBE-0542013). Author DDL was supported by NSF IGERT, R.K. Mellon Foundation, and the Program in Neural Computation (NIH Grant T90 DA022762). This research was funded in part by a grant from the Pennsylvania Department of Health's Commonwealth Universal Research Enhancement Program.

Commercial relationships: none.
Corresponding author: Daniel D. Leeds.
Email: dleeds@fordham.edu.
Address: Department of Computer and Information Science, Fordham University, Bronx, NY, USA.

References

Berg, A., Berg, T., & Malik, J. (2005). Shape matching and object recognition using low distortion correspondences. In C. Schmid, S. Soatto, & C. Tomasi (Eds.), *IEEE computer science conference on computer vision and pattern recognition, Vol. 1.* San Diego, CA: IEEE Computer Society.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 443–446.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., & Poggio, T. (2007). A model of v4 shape selectivity and invariance. *Journal of Neurophysiology*, *98*(3), 1733–1750.

Calder, A., & Young, A. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6,* 641–651.

Canny, J. (1986). A computational approach to edge detection. In David A. Forsyth (Ed.), *IEEE transactions on pattern analysis and machine intelligence, 6* (pp. 679–698). IEEE Computer Society.

Chandrasekhar, V., Makar, M., Takacs, G., Chen, D., Tsai, S. S., Cheung, N. M., Grzeszczuk, R., Reznik, Y., & Girod, B. (2010). Survey of sift compression schemes. In *Proceedings of international mobile multimedia workshop (IMMW)*, IEEE International Conference on Pattern Recognition (ICPR). Instanbul, Turkey: IEEE Computer Society.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In C. Schmid, S. Soatto, & C. Tomasi, *IEEE computer science conference on computer vision and pattern recognition, Vol. 1.* San Diego, CA: IEEE Computer Society.

Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A, 2*(7), 1160–1169.

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron, 23*(1), 115–125.

FreeSurfer. (2012). Retrieved from http://surfer.nmr.mgh.harvard.edu/

Freiwald, W., Tsao, D., & Livingstone, M. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience, 12,* 1187–1196.

Genovese, C., Lazar, N., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage, 15,* 870–878.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The FFA subserves face perception not generic within category identification. *Nature Neuroscience, 7*(5), 555–562.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41*(10–11), 1409–1422.

Harel, A., Ullman, S., Epstein, B., & Bentin, S. (2007). Mutual information of image fragments predicts categorization in humans: Neurophysiological and behavioral evidence. *Vision Research, 47*(15), 2010–2020.

Haxby, J., Hoffman, E., & Gobbini, M. (2000). The distributed human neural system for face perception. *Trends in Cognitive Science*, *4*(6), 223–233.

Haxby, J., Ungerleider, L., Clark, V., Schouten, J., Hoffman, E., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, *22*(1), 189–199.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology, 195,* 215–243.

Hummel, J., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480–517.

Hung, C., Carlson, E., & Connor, C. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, *74*(6), 1099–1113.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition. In *IEEE international conference on computer vision, (Vol. 1)*. Kyoto, Japan: IEEE Computer Society.

Just, M., Cherkassky, V., Aryal, S., & Mitchell, T. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE, 5*(1), e8622, doi:10.1371/journal.pone.0008622.

Kay, K., Naselaris, T., Prenger, R., & Gallant, J. (2008). Identifying natural images from human brain activity. *Nature, 452,* 352–355.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309.

Kimia, B., Tannenbaum, A., & Zucker, S. (1995). Shapes, shocks, and deformations I: The components of two-dimensional shape and the reaction-

diffusion space. *International Journal of Computer Vision, 15*(3), 189–224.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *The Journal of Neuroscience*, *20*(9), 3310–3318.

Kravitz, D., Peng, C., & Baker, C. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience, 31*(20), 7322–7333.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, USA, 103*(10), 3863–3868.

Kriegeskorte, N., Murr, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126–1141.

Kung, C., Peissig, J., & Tarr, M. (2007). Is region-of-interest overlap comparison a reliable measure of category specificity? *Journal of Cognitive Neuroscience, 19,* 2019–2034.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Macrini, D. (2008). Shapematcher 5. Retrieved from http://www.cs.toronto.edu/~dmac/ShapeMatcher/

MATLAB. (2012). Version 8.0.0.783 (r2012b). Natick, MA: The MathWorks Inc.

Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In A. Leonardis, H. Bischof, & A. Prinz (Eds.), *Computer vision - Ninth European Conference on Computer Vision*, Vol. 3954 (pp. 490–503). Graz, Austria: Springer-Verlag.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Olshausen, B., & Field, D. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.

O'Toole, A., Jiang, F., Abdi, H., & Haxby, J. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience*, *17*(4), 580–590.

Pelli, D. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A., & Jeeves, M. (1984). Neurones responsive to faces in the temporal cortex: Studies of functional organization, sensitivity to identity and relation to perception. *Human Neurobiology, 3*(4), 197–208.

Pittman, B. (2011). *AFNI main page — afni and nifti server for nimh/nih/phs/dhhs/usa/earth.* Retrieved from http://afni.nimh.nih.gov/afni

Pyles, J., & Grossman, E. (2009). Neural adaptation for novel objects during dynamic articulation. *Neuropsychologia*, *47*(5), 1261–1268.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Saad, Z. (2006). *SUMA - afni surface mapper – afni and nifti server for nimh/nih/phs/dhhs/usa/earth.* Retrieved from http://afni.nimh.nih.gov/afni/suma/

Schultz, J., & Pilz, K. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research*, *194*(3), 465–475.

Schyns, P., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science, 13*(5), 402–409.

Seibert, D., Leeds, D., Pyles, J., & Tarr, M. (2012). *Exploring computational models of visual object perception.* Poster presented at Vision Sciences Society Annual Meeting, Naples, Florida.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, USA, 109*(21), 6424–6429.

Siddiqi, K., Shokoufandeh, A., Dickinson, S., & Zucker, S. (1999). Shock graphs and shape matching. *International Journal of Computer Vision*, *35*(1), 13–32.

Snodgrass, J., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 174–215.

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *Neuroimage, 52*(1), 451–463.

Swisher, J., Gatenby, J., Gore, J., Wolfe, B., Moon, C.-H., Kim, S.-G., & Tong, F. (2010). Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. *Journal of Neuroscience, 30*(1), 325–330.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: Clustering of

cells with similar but slightly different stimulus selectivities. *Cerebral Cortex, 13*(1), 90–99.

Tarr, M., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3*(8), 764–769.

Tjan, B. (2000). Adaptive object representation with hierarchically-distributed memory sites. In Todd K. Leen, Thomas G. Dietterich, Volker Tresp (Eds.), *Advances in neural information processing systems 13*, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, 2000 (pp. 66–72). Cambridge, MA: MIT Press.

Torralba, A. (2006). Spatial envelope. Retrieved from http://people.csail.mit.edu/torralba/code/spatialenvelope/

Tsao, D., & Livingstone, M. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31,* 411–437.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5,* 682–687.

Vedaldi, A. (2011). A. Vedaldi -code -sift for matlab. Retrieved from http://www.vlfeat.org/~vedaldi/code/sift.html

Vogels, R. (1999). Effect of image scrambling on inferior temporal cortical responses. *Neuroreport, 10*(9), 1811–1816.

Wong, A.-N., Palmeri, T., Rogers, B., Gore, J., & Gauthier, I. (2009). Beyond shape: How you learn about objects affects how they are represented in visual cortex. *PLoS ONE, 4*(12), e8405.

Yamane, Y., Carlson, E., Bowman, K., Wang, Z., & Connor, C. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, *11*(11), 1352–1360.

Zhang, L., & Cottrell, G. (2005). Holistic processing develops because it is good. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the cognitive science society* (pp. 2428–2433). Stresa, Italy: Cognitive Science Society, Inc.