Title

Cortical networks of dynamic scene category representation in the human brain

Authors

Çelik, Emin
Keles, Umit
Kiremitçi, İbrahim
et al.

Publication Date

2021-10-01

DOI

10.1016/j.cortex.2021.07.008

HHS Public Access

Author manuscript

Cortex. Author manuscript; available in PMC 2022 October 01.

Published in final edited form as:

Cortex. 2021 October ; 143: 127–147. doi:10.1016/j.cortex.2021.07.008.

Cortical Networks of Dynamic Scene Category Representation in the Human Brain

Emin Çelik[a,b], **Umit Keles**[b,c], **İbrahim Kiremitçi**[a,b], **Jack Gallant**[d,e,f], **Tolga Çukur**[a,b,g]

[a]Neuroscience Program, Sabuncu Brain Research Center, Bilkent University, Ankara, TR-06800, Turkey

[b]National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, TR-06800, Turkey

[c]Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA-91125, USA

[d]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA

[e]Program in Bioengineering, University of California, Berkeley, CA 94720, USA

[f]Department of Psychology, University of California, Berkeley, CA 94720, USA

[g]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, TR-06800, Turkey

Abstract

Humans have an impressive ability to rapidly process global information in natural scenes to infer their category. Yet, it remains unclear whether and how scene categories observed dynamically in the natural world are represented in cerebral cortex beyond few canonical scene-selective areas. To address this question, here we examined the representation of dynamic visual scenes by recording whole-brain blood oxygenation level-dependent (BOLD) responses while subjects viewed natural movies. We fit voxelwise encoding models to estimate tuning for scene categories that reflect statistical ensembles of objects and actions in the natural world. We find that this scene-category model explains a significant portion of the response variance broadly across cerebral cortex. Cluster analysis of scene-category tuning profiles across cortex reveals nine spatially-segregated networks of brain regions consistently across subjects. These networks show heterogeneous tuning for a diverse set of dynamic scene categories related to navigation, human

Correspondence should be addressed to: Emin Çelik, National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, TR-06800, Turkey, < emin.celik@bilkent.edu.tr>.

Emin Çelik: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – Review & Editing, Visualization
Umit Keles: Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, Visualization
İbrahim Kiremitçi: Validation, Formal analysis
Jack Gallant: Writing – Review & Editing
Tolga Çukur: Conceptualization, Investigation, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition

activity, social interaction, civilization, natural environment, non-human animals, motion-energy, and texture, suggesting that the organization of scene category representation is quite complex.

Keywords

fMRI; dynamic scene category representation; voxelwise encoding model; cluster analysis

Introduction

A primary aim of visual neuroscience is to shed light on how the human brain represents diverse information in natural scenes. Behavioral research on scene perception suggests that humans categorize scenes to more efficiently process the wealth of information in visual scenes (Greene & Oliva, 2009; Konkle, Brady, Alvarez, & Oliva, 2010; Rousselet, Joubert, & Fabre-Thorpe, 2005). Therefore, it is likely that information on scene categories is represented across cortex. Consistent with this notion, previous neuroimaging studies have demonstrated that the category of a visual scene could be classified among a limited number of basic categories (e.g., beaches, forests, mountains) based on blood-oxygen level-dependent (BOLD) responses in classical scene-selective regions (parahippocampal place area, PPA; retrosplenial complex, RSC; and occipital place area, OPA), object-selective lateral occipital complex (LO), and anterior visual cortex (R. A. Epstein & Morgan, 2012; Jung, Larsen, & Walther, 2018; Walther, Caddigan, Fei-Fei, & Beck, 2009; Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). A common approach in these studies was to operationally define visual scenes into few non-overlapping categories. However, natural scene categories might show varying degrees of statistical correlation, and a real-world scene might be characterized under several distinct categories. In addition, because these studies used static scenes, they did not possess the necessary tools to demonstrate how dynamic scene categories are represented in the human brain.

To examine the statistics of natural scene categories, a recent study (Stansbury, Naselaris, & Gallant, 2013) used a data-driven algorithm to procure a broad set of scene categories wherein potential similarities between the categories were also taken into account. In this approach, each scene category is defined as a list of presence probabilities for a large array of constituent objects that appear within natural scenes. Once the algorithm learns a set of categories, the likelihood that a given scene belongs to each of the learned categories can be inferred based on the objects within the scene. This scene category model has been reported to yield improved predictions of single-voxel BOLD responses in classical face- and scene-selective areas compared to an alternative model based on the presence of a few diagnostic objects that frequently appeared in the presented natural images (Stansbury et al., 2013). This result raises the possibility that object co-occurrence statistics form the basis of scene category definitions above and beyond individual objects present in scenes.

Stansbury et al. defined categories of static scenes via their constituent objects and focused on category responses in classical scene-selective regions like many prior studies on scene representation (R. A. Epstein & Morgan, 2012; Jung et al., 2018; Walther et al., 2009, 2011). Yet, several recent studies imply that much of anterior visual cortex might be

organized by differential tuning of voxels for actions within visual scenes (Çukur, Huth, Nishimoto, & Gallant, 2016; Tarhan & Konkle, 2020). In fact, real-world scenes contain dynamic interactions between objects and actions leading to more elaborate categories (Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016), and they have been reported to elicit widespread responses across visual cortex (Deen, Koldewyn, Kanwisher, & Saxe, 2015; R. A. Epstein & Baker, 2019; Isik, Koldewyn, Beeler, & Kanwisher, 2017; Maguire et al., 1998). Therefore, it is likely that natural scene categories based on co-occurrence of objects and actions are represented across broadly distributed networks in the human brain.

Here, we sought to learn high-level features that capture scene-category information in dynamic visual scenes, and to examine how this information is represented across cerebral cortex. We first recorded BOLD responses while subjects viewed a large set of natural movies that contained 5252 distinct objects and actions. To identify scene-category features, we employed a statistical learning algorithm that learned a large set of categories on the basis of the co-occurrence statistics of objects and actions in the natural world. We then used the learned scene categories within a voxelwise modeling framework (Çukur et al., 2016; Çukur, Nishimoto, Huth, & Gallant, 2013; Huth, Nishimoto, Vu, & Gallant, 2012; Nishimoto et al., 2011) to estimate scene-category tuning profiles in single voxels across cerebral cortex. Subsequently, we performed a clustering analysis in order to reveal large-scale networks of brain regions that differ in their scene-category tuning.

Materials and Methods

Experimental Design

Subjects.—Five healthy human subjects (all male, ages 25–32 years) with normal or corrected-to-normal vision participated in this study. MRI data were collected in five separate scan sessions: three sessions for the main experiment, one session for acquiring functional localizers, and one session for acquiring anatomical data. Experimental protocols were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley. All subjects gave written informed consent prior to scanning.

MRI protocols.—MRI data were collected on a 3T Siemens Tim Trio scanner (with a 32-channel head coil) located in the Brain Imaging Center at the University of California, Berkeley. T2*-weighted functional data were acquired using a gradient-echo echo-planar imaging sequence with the following parameters: repetition time (TR) = 2 s, echo time (TE) = 31 ms, a water-excitation pulse with flip angle = 70°, voxel size = 2.24 × 2.24 × 3.5 mm^3, field of view = 224 × 224 mm^2, and 32 axial slices spanning across the entire brain. Anatomical scans were performed with a T1-weighted magnetization-prepared rapid-acquisition gradient-echo sequence with the following parameters: TR = 2.30 s, TE = 3.45 ms, flip angle = 10°, voxel size = 1 × 1 × 1 mm^3 and field of view = 256 × 256 × 192 mm^3.

Main experiment.—Whole-brain BOLD responses were recorded while subjects passively viewed two hours of color natural movies. The movies were compiled by combining 10–20 s movie clips selected from the Apple QuickTime HD gallery and YouTube.com (Nishimoto et al., 2011). The stimulus was separated into two independent sets. The first set was used

to train voxelwise encoding models and it consisted of 12 separate runs of 10 min each. Each clip appeared only once in the training set. The second set was used to test the performance of fit models, and it consisted of 9 runs of 10 min each. Each 10 min run was composed of the same ten 1 min blocks, but the presentation order of the blocks were randomly shuffled in each run. Each 1 min block was presented 9 times in total and the respective BOLD responses were averaged across repeats. The movies (512 × 512 pixels) were presented at a 24° × 24° visual angle, using an MRI-compatible projector (Avotec) and a custom-built mirror system. A fixation spot (0.16° × 0.16°) with alternating color (3 Hz) was overlaid onto the movies. Note that the dataset reported here was also analyzed in several other studies (Çelik, Dar, Yılmaz, Keleş, & Çukur, 2019; Çukur et al., 2016; Çukur, Huth, Nishimoto, & Gallant, 2013; Huth et al., 2012). The experimental stimuli are available at https://crcns.org/data-sets/vc/vim-2. Subject data and labeled stimuli are available at https://crcns.org/data-sets/vc/vim-4.

Functional localizers.—Functional localizer and retinotopic mapping data were acquired separately from the main experiment. Localizers for category-selective regions of interest (ROIs) were acquired in six 4.5 min runs, each divided into 16 blocks. Each block lasted 16 s and contained 20 static images from each of the following categories: human faces, human body parts, non-human animals, household objects, spatially scrambled objects, and places. These category blocks were displayed in a different order in each run. Each image was displayed for 300 ms, and 500 ms blank periods were inserted between consecutive images. To sustain vigilance, subjects were instructed to press a button when two consecutive images were identical. The localizer for area V5/MT+ was acquired in four 90 s runs, each divided into 6 blocks. Each block contained 15 s of continuous or temporally scrambled natural movies (Tootell et al., 1995). Retinotopic mapping data were acquired in four 9 min runs. The runs contained clockwise or counter-clockwise rotating polar wedges, and expanding or contracting rings, respectively (Hansen, Kay, & Gallant, 2007).

Data preprocessing.—The FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (Jenkinson, Bannister, Brady, & Smith, 2002) was used for functional alignment. First, intra-run transformations were estimated to align volumes within each run. A template volume was then generated for each run as the temporal average of the aligned volumes. In each subject, the template volume of the first run in the first experimental session was selected as the target template. Afterwards, inter-run transformations were estimated between the template of each run and the target template. The transformations for intra-run and inter-run alignment were combined, and finally applied on fMRI data in a single step. Following alignment, low-frequency drifts in BOLD responses were removed from each voxel using a median filter with a 120 s temporal window. Lastly, each voxel's response was normalized to zero mean and unit variance across time. No temporal or spatial smoothing was applied to the functional data collected in the main experiment. The functional localizer data were also motion-corrected and aligned to the target template of the main experiment. The localizer data were smoothed using a Gaussian kernel of 4-mm full-width at half-maximum (Spiridon, Fischl, & Kanwisher, 2006).

Definition of functional ROIs.—Standard procedures (Spiridon et al., 2006) were used to define functional ROIs in each subject. Functional localizer data were examined to identify contiguous groups of voxels that showed significantly stronger responses to a specific stimulus category according to standard contrasts (t test, $p < 10^{-4}$, uncorrected). Places versus isolated objects contrast was used to define the PPA in the parahippocampal gyrus (Aguirre & D'Esposito, 1997; R. Epstein & Kanwisher, 1998), the RSC in the retrosplenial sulcus (Maguire, 2001), and the OPA in the temporal-occipital sulcus (Dilks, Julian, Paunov, & Kanwisher, 2013). In addition, faces versus objects contrast was used to define the fusiform face area (FFA) (Kanwisher, McDermott, & Chun, 1997) and occipital face area (OFA) (Gauthier, Tarr, et al., 2000). Human body parts versus objects contrast was used to define the extrastriate body area (EBA) (Downing, Jiang, Shuman, & Kanwisher, 2001). Objects versus spatially scrambled objects contrast was used to define the lateral occipital complex (LO) (Malach et al., 1995). Continuous versus temporally scrambled movies contrast was used to define the area V5/MT+ (Tootell et al., 1995). Last, the retinotopic mapping data were used to define the early visual areas (V1–4) following standard procedures (Engel, Glover, & Wandell, 1997; Hansen et al., 2007).

Visualization on flatmaps.—The organization of scene category representation across cortex was visualized by using flattened cortical maps. Individual subjects' flatmaps were generated from their anatomical data (T1-weighted brain scans) using Caret5 software (Van Essen et al., 2001). Information about the cluster memberships of individual voxels was projected onto the flattened cortical maps by aligning functional and anatomical data, using the Pycortex package (Gao, Huth, Lescroart, & Gallant, 2015).

Voxelwise Scene Category Models

An encoding model was used to measure voxelwise tuning for scene categories (Figure 1). In addition, a control model based on parts of scenes was used to measure voxelwise tuning for object and action components of natural scenes. The performances of these two models were compared in terms of the variance they explained in recorded BOLD responses.

Scene-category model.—A comprehensive model of scene categories that build on constituent objects and actions in scenes is lacking. Thus, to objectively identify the features of the scene-category model, we used a data-driven approach based on the latent Dirichlet allocation (LDA) algorithm (Blei, Ng, & Jordan, 2003; Phan & Nguyen, 2007). LDA was originally proposed to uncover latent topics from a large text corpus on the basis of word co-occurrence statistics (Blei et al., 2003). A recent study (Stansbury et al., 2013) used LDA to learn scene categories from a database of natural images based on object co-occurrence statistics. Here, we used the LDA algorithm to learn dynamic scene categories that capture co-occurrence statistics of not only objects, but also actions in dynamic natural scenes.

LDA was performed on a large training corpus of movie scripts and scene annotations containing 5252 distinct object and action categories (see Training corpus for details). Scene-category features were learned that capture the co-occurrence statistics of objects and actions in this corpus. Each scene-category feature was a 5252-dimensional vector that reflected the probability of occurrence for individual objects and actions within the

respective scene category. Given the list of objects and actions in a novel scene, an LDA-based inference procedure can also be performed to calculate the probability that the scene belongs to a particular scene category. This procedure was performed on each 1-s clip of the movies, yielding for each clip a probability distribution over the scene-category features.

Representative scene-category features are shown in Figure 1. The features correspond to natural scene categories that capture the co-occurrence of multiple objects and actions frequently encountered in the real world. For example, one feature reflects information about urban street scenes: “in a street view, a bus is driven on a road while a truck, a park appears in the background” (C1 in Figure 1). Another feature reflects information about a roadway scene: “a car is driven by a driver on a road” (C2 in Figure 1).

Training corpus.—A training corpus was compiled to learn the features of the scene-category and part-of-scene models. This corpus consisted of the annotations in the Microsoft COCO data set (Lin et al., 2014), the Microsoft Research video description corpus (Chen & Dolan, 2011), and subtitles of 4068 documentaries. Raw text in the compiled corpus contained 26 million words in approximately 700,000 separate entries. Standard preprocessing routines were applied including tokenization, stemming, removal of frequent stop words, and part-of-speech tagging (Bird, Klein, & Loper, 2009) to only retain objects (i.e., nouns) and actions (i.e., verbs). Following preprocessing, the corpus contained 10 million words with a vocabulary of nearly 28,000 words. The vocabulary was further reduced to 5252 objects and actions that commonly appeared in both the corpus and the movie descriptions provided by Amazon Mechanical Turk workers.

Stimulus time courses.—To project the movies onto the features of the scene-category model, objects and actions that appeared in each 1-s clip were manually labeled. During labeling, the WordNet lexicon (Miller, 1995) was used to take into account hierarchical relationships among object and action categories (Huth et al., 2012). For instance, for a clip containing the object “dog”, labels for superordinate categories “canine”, “carnivore”, “mammal”, “animal”, “organism”, and “living thing” were also assigned. Next, the LDA algorithm was used to infer the distribution of model features in each 1-s movie clip based on the constituent object and action labels. The distributions were aggregated across clips to form the stimulus time course.

High-level semantic features in natural visual stimuli may be partly correlated with low-level motion-energy features, including spatiotemporal frequency, spatial position, and orientation (Çukur et al., 2016; Çukur, Nishimoto, et al., 2013; Lescroart, Stansbury, & Gallant, 2015). To reduce spurious correlations, a motion-energy regressor was appended to the scene-category model. The motion-energy features were previously shown to account for BOLD responses elicited by natural movies in early visual areas (Nishimoto et al., 2011). To calculate the motion-energy features of the movies, the movie frames were filtered with 2,139 spatiotemporal Gabor wavelets at 3 temporal frequencies (0, 2, and 4 Hz), 6 spatial frequencies (0, 1.5, 3, 6, 12, and 24 cycles/image) and 8 directions (0, 45, 90, …, 315 degrees). The nuisance regressor characterized the total motion energy as the summed output of 2,139 Gabor filters.

Statistical Analysis

Model fitting.—A voxelwise modeling framework was used to measure single-voxel tuning in individual subjects (Çukur et al., 2016; Çukur, Huth, et al., 2013; Çukur, Nishimoto, et al., 2013; Ester, Sprague, & Serences, 2020; Han et al., 2019; Huth et al., 2012; Serences & Saproo, 2012; Wen et al., 2017). In this framework, each voxelwise model contains a set of weights that reflect the effect of individual model features on the voxel's responses. Because model weights are signed, they do not only capture response magnitude but also differentiate between relative increases and decreases in responses. For the scene-category model, the direction in which a given scene-category feature effects BOLD responses will therefore be captured in the sign of the corresponding model weight. Note that subsequent clustering analyses are also based on fit model weights. As such, the differentiation between characteristic response increases/decreases will be reflected in functional cluster definitions. Models were fit to optimally predict measured BOLD responses (Figure 1, see https://github.com/icon-lab/SPIN-VM). Fitting was performed on 7200 s of training data using linear regression with l_2-regularization. The stimulus time course was temporally down-sampled to match the temporal sampling rate of BOLD responses. To capture hemodynamic delays, separate finite-impulse-response (FIR) filters were appended to each feature. These filters introduced temporal delays of two, three, and four samples (or equivalently 4, 6, and 8 s). The FIR coefficients were fit together with the model weights. A 20-fold cross validation procedure was used to determine the optimal regularization parameter (λ) for each voxel. In each fold, voxelwise models were trained on a randomly sub-sampled set (at a rate of 90%) of the training data. Model performance was then measured on the held-out 10% of the training data. Performance was taken as the correlation coefficient (Pearson's r) between the measured and predicted BOLD responses. Raw correlation coefficients are biased downward by noise in the measured BOLD responses (David & Gallant, 2005). Hence, correlation coefficients were corrected for noise bias following the procedure detailed in (Huth et al., 2012). The optimal λ for each voxel was selected to maximize average model performance across cross-validation folds. Voxelwise models were refit using this optimal λ on the entire training data. The performance of the fit models was then evaluated on 540 s of test data using a 10,000-fold jackknife resampling (at a rate of 80%) procedure. Prediction scores were measured on jackknife samples. Significance level was taken as the proportion of jackknife samples with negative scores. Corrections for multiple comparisons were conducted using false-discovery-rate (FDR) control (Benjamini & Yekutieli, 2001).

Control Models

Part-of-scene model.—To identify object-action components of natural scenes, we used a data-driven approach based on the non-negative matrix factorization (NMF) algorithm (Lee & Seung, 1999; Pedregosa et al., 2011). An original application of the NMF algorithm was extraction of sparsely distributed, additive semantic features from a large text corpus (Lee & Seung, 1999). Here, NMF was performed on the same training corpus as LDA to learn part-of-scene features that reflect object and action components of natural scenes (Supp. Figure 1). Each part-of-scene feature was a 5252-dimensional vector that reflected the probability of occurrence for individual object and action categories. Given a novel

scene, an NMF-based inference procedure was performed to express the distribution of part-of-scene features within that scene. This procedure was performed on each 1-s clip of the movies, yielding for each clip a distribution over the part-of-scene features. Representative part-of-scene features are shown in Figure 2.

Gist model.—We used voxelwise gist models to measure tuning for low-level spatial features in the movies. While early visual areas are commonly assumed to represent low-level features of visual scenes (Grill-Spector & Malach, 2004), it remains unclear what specific information in natural scenes is represented in PPA, OPA, and RSC (R. A. Epstein, 2014). Recent studies suggest that these areas might represent both low-level spatial features (Kravitz, Peng, & Baker, 2011; Park, Brady, Greene, & Oliva, 2011; Watson, Hartley, & Andrews, 2014), and high-level semantic features (Çukur et al., 2016; Huth et al., 2012; Stansbury et al., 2013; Walther et al., 2009).

Voxelwise gist models were fit to measure tuning for spatial layout and texture of visual scenes. The gist features were shown to be effective in capturing the global spatial properties such as openness, expansion, and texture of natural scenes (Oliva & Torralba, 2001). To calculate the gist features of the movies, the movie frames were first down-sampled to 256 × 256 pixels. A total of 512 gist features were then computed in 16 image blocks, each containing 4 spatial scales and 8 orientations per scale (Oliva & Torralba, 2001).

Model selection.—Humans can perceive a vast number of scene categories as well as constituent object and action categories within scenes (Çukur et al., 2016; Huth et al., 2012; Stansbury et al., 2013). However, because the spatiotemporal resolution of fMRI is coarse, BOLD responses will admit sensitive examination of only a portion of these categories (Stansbury et al., 2013). Thus, we separately identified the set of scene-category and the set of part-of-scene features that best explain measured BOLD responses across subjects.

To do so, we incremented the number of features learned by the LDA (for the scene category model) and NMF (for the part-of-scene model) algorithms from 10 to 200. We fit separate voxelwise models for each distinct number of features, and we measured the relative number of significantly predicted cortical voxels across subjects. We find that the optimal number of features is 180 for the scene-category model and 190 for the part-of-scene model (Figure 3). Because the optimal numbers of features were very close across models, the scene-category model with 180 features and the part-of-scene model with 190 features (as control model) were used in subsequent analyses.

We also note that although the best performance for the scene-category model is attained with 180 features, a reduced model based on 60 features has a close performance. The difference between these two models in terms of the number of significantly predicted cortical voxels was less than 0.02% across subjects. Yet the smaller number of features in the reduced model offers an advantage in visualization and interpretation of scene category representations. Thus, we used this reduced model in cluster analysis and subsequent visualization on the cortical surface.

Cluster Analysis.—The natural movies used here span a broad variety of complex real-world scenes. The movies contain static scenes involving objects such as urban views or landscapes and complex, dynamic scenes that involve both objects and actions such as locomotion or social interaction. A core issue that this report addresses is how these various scene categories are represented in the brain. To investigate this issue, we performed cluster analysis on voxelwise tuning profiles (i.e., vectors of model weights) for scene categories. Because the specific areas that are involved in scene category representation remain unclear, the cluster analysis included all cortical voxels significantly predicted by the scene-category model ($p < 0.05$, FDR corrected). Clusters were obtained via the k-means algorithm, where similarity of voxelwise tuning profiles was measured using Euclidean distance (James, Witten, Hastie, & Tibshirani, 2013). To avoid unstable clustering solutions, we employed k-means++ with smart initialization of cluster centers (Arthur & Vassilvitskii, 2007). Lastly, each cluster center was taken as the mean tuning profile across voxels within that cluster.

To examine the consistency of clusters across subjects, we performed a cluster analysis on each subject separately and obtained individual-subject cluster centers. Only the significantly predicted voxels ($p < 0.05$, FDR corrected) in each subject were included in individual-subject level analyses. We also performed a cluster analysis after pooling voxels across subjects and obtained group-level cluster centers. To prevent bias due to across-subject variability in signal-to-noise ratios of fMRI data and brain sizes, a fixed number of voxels were selected from each individual subject in the group-level analysis. A minimum of 10222 significantly predicted voxels were obtained for subject S3 ($p < 0.05$, FDR corrected), therefore a total of 51110 voxels were included across all five subjects. Across-subject consistency was then assessed by measuring the similarities between individual-subject cluster centers and also by measuring the similarities between individual-subject cluster centers and the group-level cluster centers. Similarity was taken as the correlation coefficient (Pearson's r) between the cluster centers.

A critical hyperparameter for cluster analysis is the number of voxel clusters to recover. We determined the number of voxel clusters using an unsupervised procedure. This procedure measured the proportion of explained variance in tuning profiles by the respective cluster centers. To do this, the total variance in tuning profiles was measured across all voxels. Then, within-cluster variances in tuning profiles across voxels were identified within each cluster. Subsequently, the difference between the total variance and the sum of within-cluster variances across clusters was calculated. The proportion of explained variance in tuning profiles was taken as the ratio of this difference to the total variance (James et al., 2013). The optimal number of clusters was selected as the number beyond which the improvement in explained variance fell below one percent, since at that point clusters started to differentiate between subjects as opposed to functional selectivity profiles.

Scene categories that elicit differential responses across voxel clusters were manually labeled (see Figure 4). Labeling for each scene category was performed by visual inspection of the top five movie frames that yielded the maximum probability for that category. To ensure reliability of the scene category labels, four healthy adult males who were naïve as to the purposes of the study were asked to rate the labels. A 5-point Likert scale was used to measure labeling accuracy. The raters were asked to inspect the movie frames for each scene

category, rate the assigned labels for accuracy (1 = inaccurate, 3 = moderately accurate, 5 = accurate), and provide their own labels. Rater consistency was measured by pooling all ratings and calculating the average and standard error of the mean (SEM). Cluster labeling was also based on the same four raters' suggestions by majority voting.

Control for stimulus sampling bias.—Separate scene categories in natural movies may contain a shared subset of objects or actions. For instance, the object "human body" and the action "jumping" can take part in "a concert" scene as well as in "a sports activity" scene. Thus, the scene-category features learned here might show correlations in terms of their distribution over objects and actions. In turn, these correlations may bias the voxelwise tuning profiles for scene categories and subsequent cluster analysis on these profiles. To assess whether our results are biased by this potential confound, we performed an additional cluster analysis based on the stimulus time course of scene-category features. To control for temporally lagged stimulus correlations, we generated multiple time courses for scene-category features temporally delayed by lags from −5 to 5 s. We averaged these time courses to obtain an aggregate stimulus matrix (time × scene-category features). We then performed cluster analysis on the aggregate stimulus matrix to group movie clips into clusters based on their scene category distributions. Stimulus cluster centers were taken as the mean profile of scene-category features across movie clips within each cluster. Finally, we compared the variance explained in voxelwise tuning profiles by the voxel cluster centers to that explained by the stimulus cluster centers. For this comparison, each voxel was assigned to a stimulus cluster center that was most similar to its tuning profile.

Power Analyses

Several a priori power analyses were conducted for statistical assessment of model prediction scores. As in the main analyses, prediction score was taken as the correlation coefficient between measured and predicted BOLD responses. First, to determine the minimum detectable effect size for single-voxel prediction scores, a Monte Carlo procedure of 1000 iterations was performed. During each iteration, "measured" and "predicted" BOLD responses were simulated as sets of random samples from a bivariate normal distribution. The set size was taken as 270 to match that of the test set used in the main analyses. Both variables in the normal distribution had zero mean and unit variance. The effect size was systematically controlled by varying the correlation between the two variables. Given a measured-predicted response set, responses for a single voxel were resampled without replacement using a 10,000-fold jackknife resampling (at a rate of 80%) procedure to calculate significance level (p). Power was taken as the percentage of Monte Carlo iterations with significant test results ($p < 0.05$). For a desired power level of 0.8, the minimum detectable effect size in single-voxel prediction scores was 0.05.

Second, a Monte Carlo simulation was performed to detect the minimum detectable effect size for differences in single-voxel prediction scores between two competing models. In this case, two distinct bivariate normal distributions were used to simulate prediction scores from the two models, respectively. For the first model taken as a reference, the correlation between the variables in the bivariate distribution was set to zero. For the second model, the correlation between the variables in the bivariate distribution was systematically varied.

Given a pair of measured-predicted response sets from the two models, responses were again resampled without replacement 10,000 times at a rate of 80% to calculate significance level (p). Power was taken as the percentage of Monte Carlo iterations with significant test results ($p < 0.05$). For a desired power level of 0.8, the minimum detectable effect size in difference of single-voxel prediction score between two models was 0.05.

Lastly, a Monte Carlo simulation was performed to detect the minimum detectable effect size in ROI-level prediction scores between competing models. Note that the smallest ROI examined in this study contained more than 10 voxels. Therefore, simulations were run for a hypothetical ROI with 10 voxels. The simulations for between-model differences in single-voxel prediction score were expanded to include 10 independent voxels. In each iteration, significance level was calculated after averaging prediction scores across 10 voxels within the ROI. Power was taken as the percentage of Monte Carlo iterations with significant test results ($p < 0.05$). For a desired power level of 0.8, the minimum detectable effect size in ROI-level difference in prediction score between two models was 0.02.

A separate power analysis was conducted for statistical assessment of cluster-center correlations across subjects. As in the main analyses, similarity was taken as the correlation coefficient between the cluster centers. To determine the minimum detectable effect size, a Monte Carlo procedure of 1000 iterations was performed. During each iteration, cluster centers were simulated as sets of random samples from a bivariate normal distribution. The number of clusters was taken as 9 to match the number of clusters used in the main analyses. Both variables in the normal distribution had zero mean and unit variance. The effect size was systematically controlled by varying the correlation between the two variables. Cluster centers were resampled without replacement using a 10,000-fold bootstrap resampling procedure to calculate significance level (p). Power was taken as the percentage of Monte Carlo iterations with significant test results ($p < 0.05$). For a desired power level of 0.8, the minimum detectable effect size in cluster center correlations was 0.055. Effects measured in the main analyses were only deemed significant if they exceeded the minimum detectable effect sizes identified by these a priori power analyses.

Results

To investigate the nature of high-level scene information that is represented across the cerebral cortex, we recorded BOLD responses while subjects passively viewed 2 h of natural movies. We used voxelwise modeling to assess scene representations in single voxels. We fit a scene-category model to measure tuning for scene categories (e.g., an urban street, a forest) that reflect co-occurrence statistics of objects and actions in natural scenes. Model performance was evaluated by calculating voxelwise prediction scores on BOLD responses reserved for this purpose. Prediction scores were assessed in PPA, OPA, and RSC, as well as several other classical functional ROIs including intraparietal sulcus (IPS), posterior superior temporal sulcus (pSTS), fusiform face area (FFA), extrastriate body area (EBA), human MT (V5/MT+), lateral occipital complex (LO), visual retinotopic area V7, and early visual areas (RET: V1–V3). We also fit a part-of-scene model that acted as a control model. This model measures tuning for scene components (e.g., a car, a road, or a driving action) that reflect constituent objects and actions in natural scenes (see Supp. Figure 1).

The scene-category and part-of-scene models significantly predict BOLD responses both at the group- and the individual-subject levels (jackknife test; $p < 0.05$ (FDR corrected), prediction score (r) > 0.05). Yet, we find that the scene-category model outperforms the control model in single voxels distributed across much of cortex (jackknife test; $p < 0.05$, $\Delta r > 0.05$). Among voxels that are significantly predicted by either of the two models, the proportion of voxels in which the scene-category model outperforms the control model is given in Table 1 ($\Delta r > 0.02$; see Supp. Table 1 for the proportion of voxels where the control model outperforms the scene-category model). The scene-category model shows dominant performance in PPA, OPA, RSC, IPS implicated in spatial attention, and pSTS implicated in representing human-object interactions. In contrast, the control model appears relatively dominant, albeit to a lesser degree, in face-selective FFA and object-selective V7. Meanwhile, the two models show relatively balanced performance in RET, LO, MT+, and EBA. Thus, scene-category representations are more dominant in brain regions involved in various aspects of scene processing, and scene-category and object representations are equally dominant in many other visual areas except some specialized object-selective regions. Overall, these results suggest that many cortical regions represent holistic information about scene categories beyond information conveyed by constituent object and action components. Therefore, functional selectivity as measured by the scene-category model was further examined in subsequent analyses to assess scene category representations across cortex.

Organization of Scene Category Representation Across the Cerebral Cortex

Several previous studies provided evidence that category representations are organized into a multi-dimensional semantic space distributed systematically across the cerebral cortex (Haxby et al., 2011; Huth et al., 2012). We have recently shown that this organization is apparent even within classical category-selective areas, such as FFA and PPA, resulting in several spatially-segregated functional voxel clusters with distinct semantic tuning profiles in each ROI (Çukur et al., 2016; Çukur, Huth, et al., 2013). Collectively, these results imply that representation of scene category information shows a more fine-grained cortical organization than typically assumed (Grill-Spector & Weiner, 2014). Therefore, we hypothesized that scene categories that reflect the co-occurrence statistics of objects and actions are systematically represented in multiple spatially-segregated functional voxel clusters across the cerebral cortex.

To examine the cortical organization of scene category representation, we performed cluster analysis on voxelwise tuning profiles that were estimated by the scene-category model. We first determined the optimal number of voxel clusters by examining the variance in tuning profiles that was explained by cluster centers (see Materials and Methods). The optimal number of voxel clusters was determined as nine (see Supp. Figure 2). These nine clusters were identified by pooling voxels across subjects. But it was not clear how similarly these clusters were expressed in individual subjects. Thus, we examined the inter-subject consistency of cluster centers by assessing the correlation coefficient between individual-subject cluster centers. We find that the individual-subject cluster centers are significantly correlated across subjects ($r = 0.52 \pm 0.02$, mean ± sem across subjects; bootstrap test, $p < 0.05$) and they are significantly correlated with the group cluster centers ($r = 0.70 \pm 0.03$;

$p < 0.05$). These results suggest that scene category representation is organized into nine functional voxel clusters consistently in each individual subject. Because the cluster centers are highly consistent across subjects, to facilitate inter-subject comparisons and enhance sensitivity, we used the group-level clustering in subsequent analyses.

We expected these 9 voxel clusters to be tuned to different scene categories and that there would be a meaningful pattern providing a sensible functional interpretation. To determine these differences, we measured the average response of each cluster to 20 scene categories frequently observed in daily life (Figure 4). These 20 categories were labeled by the authors and rated by four healthy adult males (non-authors). Rater consistency was very high (4.71 ± 0.06; 1 = inaccurate labeling, 5 = accurate labeling, lowest category = 3.75 ± 0.48) across categories. Figure 5 shows the average responses of each cluster. Clusters were named based on the same four raters' suggestions. Cluster 1 contains $12\% \pm 2\%$ (mean ± sem across subjects) of the significantly predicted voxels across subjects on average, and yields greater responses to scenes showing navigation, such as a car driven on a road or a pedestrian walking, and in a lesser degree, to scenes showing a mountain or a seaside. Reduced responses are observed for scenes depicting verbal communication, which mostly contain close-up views of human faces ($p < 0.05$, bootstrap test).

Clusters 2–4 contain $9\% \pm 2\%$, $11\% \pm 2\%$, and $8\% \pm 2\%$ of the significantly predicted voxels across subjects on average, respectively. These yield greater responses to scenes showing humans and human-made environments and artifacts, such as a person jumping, cycling or walking outdoors, and humans engaged in verbal communication ($p < 0.05$). Reduced responses are observed for scenes that contain landscapes such as a mountain or a seaside view ($p < 0.05$). More specifically, Cluster 3 yields greater responses to scenes depicting social communication, such as verbal or textual communication ($p < 0.05$) and Cluster 4 yields greater responses to urban scenes with human-made environments and artifacts, such as vehicles, train stations, or trains ($p < 0.05$). For these two clusters, reduced responses are observed for scenes showing natural environments or non-human animals ($p < 0.05$).

Cluster 5 contains $15\% \pm 3\%$ of the significantly predicted voxels across subjects on average, and yields greater responses to scenes showing natural environments such as a mountain, a body of water, or aquatic animals ($p < 0.05$). Cluster 6 contains $7\% \pm 2\%$ of the significantly predicted voxels across subjects on average, and yields greater responses to scenes showing non-human animals ($p < 0.05$). For these two clusters, reduced responses are observed for scenes showing humans and human-made environments ($p < 0.05$).

Clusters 7 and 8 contain $12\% \pm 2\%$ and $13\% \pm 3\%$ of the significantly predicted voxels across subjects on average, respectively. These yield greater responses to low-level features of the movies. More specifically, Cluster 7 yields greater responses to motion-energy in the movies, such as a scene showing a person involved in a sports activity or in locomotion; or an animal or a vehicle in motion. Reduced responses are observed for static scenes such as a mountain view ($p < 0.05$). Cluster 8 yields greater responses to texture in scenes such as dynamic text on a smooth background or a flying object against a cluttered background (p

< 0.05). Finally, Cluster 9 contains 13% ± 1% of the significantly predicted voxels across subjects on average, but these voxels are not selective for any particular scene category.

Regions of fMRI signal dropout are shown in dark gray. Conventional ROIs identified using separate functional localizers are labeled and marked with white boundaries. Major anatomical landmarks and sulci are also shown: central sulcus, CeS; cingulate sulcus, CiS; collateral sulcus, CoS; inferior frontal sulcus, IFS; intraparietal sulcus, IPS; inferior temporal sulcus, ITS; middle temporal sulcus, MTS; prefrontal cortex, PfC; postcentral sulcus, PoCeS; precuneus, PrCu; superior frontal sulcus, SFS; superior temporal sulcus, STS; temporo-parietal junction, TPJ.

Cortical Maps of Scene Category Representation

To visualize the distribution of scene category tuning across cerebral cortex, we projected the clusters onto the cortical flatmaps of individual subjects (Figure 6). The distribution of voxel clusters across the cerebral cortex reveals that scene category representation is organized in nine spatially-segregated networks of brain regions. We named the networks according to their respective scene-category tuning as revealed by inspection: navigation, human activity, social interaction, civilization, natural environment, non-human animal, motion-energy, texture, and low-category selectivity networks.

Navigation: This network shows high selectivity for navigational scenes and is distributed broadly across occipital, posterior parietal, and ventral temporal cortices. The navigation network overlaps with scene-selective areas PPA, OPA, and RSC. It also includes voxels located in posterior subregions of IPS and some voxels located near superior and posterior primary somatosensory cortex (S1F), regions that have been associated with visual attention (Culham & Kanwisher, 2001; Posner, Sheese, Odludaş, & Tang, 2006).

Human Activity: Networks related to animacy, such as human activity, social interaction, civilization, and non-human animals, are distributed broadly across occipital, posterior parietal, and ventral temporal cortices. The human activity network includes several previously identified functional areas in occipitotemporal cortex that represent human faces and bodies, such as FFA, OFA, and EBA (Kanwisher, 2010). The human activity network also overlaps with MT+ and the posterior bank of the inferior temporal sulcus (ITS), two areas suggested to be involved in processing of biological motion (Thompson, Clarke, Stewart, & Puce, 2005).

Social Interaction: The social interaction network includes areas that have previously been associated with processing of social information and theory of mind (Saxe, 2006). More specifically, within temporal cortex, this network includes voxels located in pSTS, an area previously linked to face perception, human motion and actions, and social interaction (Deen et al., 2015; Isik et al., 2017). The social interaction network also includes voxels in parietal cortex that run along the anterior regions of precuneus (PrCu), an area previously associated with social cognition (Cavanna & Trimble, 2006). Finally, it also includes voxels in frontal cortex that are located in the inferior frontal sulcus face patch (IFSFP) which

has previously been linked to visual speech processing (Calvert & Campbell, 2003; Tsao, Moeller, & Freiwald, 2008).

Civilization: The civilization network shows high selectivity for human-made artifacts and environments. It neighbors the social interaction network and includes voxels near pSTS, which has previously been linked to the representation of actions involving human-made objects (Kable & Chatterjee, 2006).

Non-human Animal: The non-human animal network is broadly distributed across several ventral temporal areas near FFA and EBA, and includes voxels that are located in V7 and V3B. The non-human animal network also includes voxels located in the postcentral sulcus (PoCeSu). These regions have previously been associated with the representation of non-human animals (Huth et al., 2012; Mruczek, Von Loga, & Kastner, 2013).

Others: The remaining networks that predominantly represent motion and form information in visual scenes are broadly distributed across striate, extrastriate, occipitotemporal, and parietal cortices. The natural environment network includes voxels within occipital cortex, mainly in retinotopically organized early visual areas (V1-V4). The texture network also largely overlaps with retinotopically organized early visual areas, as expected. Lastly, the motion-energy network includes voxels located in V3A, V7, EBA, and MT+, as well as anterior IPS and superior bank of PoCeSu.

Distribution of Networks within Conventional Functional ROIs

Previous neuroimaging studies have suggested several brain areas in ventral temporal cortex that are homogeneously tuned for specific categories, such as face-selective area FFA and scene-selective areas PPA, OPA, and RSC (Kanwisher, 2010). However, recent studies have indicated that these classical ROIs consist of several functional subdomains with differential tuning for individual object and action categories (Çukur et al., 2016; Çukur, Huth, et al., 2013; Weiner, Sayres, Vinberg, & Grill-Spector, 2010). Thus, it is possible that some ROIs might contain distinct functional subdomains that exhibit differential tuning for scene categories. To test this functional heterogeneity hypothesis, we examined the relative size of the nine scene-category networks within functional ROIs. Specifically, we measured the percentage of voxels that belong to each network in PPA, OPA, RSC, FFA, EBA, MT+, pSTS, LO, V7, and retinotopically organized early visual areas (Figure 7).

First, we examined the proportion of networks within scene-selective areas PPA, OPA, and RSC. We find that the navigation network is dominant in PPA (74% ± 6%, mean ± sem across subjects, $p < 0.05$, bootstrap test, FDR corrected), OPA (70% ± 6%, $p < 0.05$), and RSC (61% ± 10%, $p < 0.05$). This result is consistent with previous reports suggesting that voxels in these areas respond selectively to stimuli that contain scenes depicting navigation (R. A. Epstein, 2008). In addition to the navigation network, there is a considerable proportion of voxels in RSC (26% ± 8%, $p < 0.05$) that respond selectively to stimuli that fall into the civilization network. Several previous reports suggest that RSC might play a different role in representation of scenes compared to PPA and OPA (Malcolm, Groen, & Baker, 2016). Whereas PPA and OPA are assumed to represent local scene

information, RSC is hypothesized to represent the broader environment likely to capture navigationally-relevant information in the surroundings (R. A. Epstein, 2008). Because the presence and locomotion of humans can serve navigational cues in the real world, the larger portion of civilization network in RSC might be a reflection of the distinct functional role of RSC.

Functional Heterogeneity in Scene Selective Areas

We recently provided evidence that PPA, OPA, and RSC each contain two functional subdivisions that differ in their responses to static scenes that show human-made artifacts such as buildings, furniture, instruments versus dynamic scenes that show human and animal locomotion and vehicles (Çukur et al., 2016). That previous study aimed to examine the organization of object and action category representation within scene-selective areas, whereas the current study examines the large-scale organization of scene category representation across the cerebral cortex. Therefore, we hypothesized that there would be two distinct functional subdomains in PPA, OPA, and RSC. To test this hypothesis, we performed an additional ROI-wise cluster analysis based on voxelwise scene-category tuning profiles (not shown). We find that ROI-wise cluster analysis identifies two functional subdomains in PPA, RSC, and OPA. The first subdomain is tuned for static scene categories such as urban or natural environments, and the second subdomain is tuned for dynamic scene categories such as human and animal locomotion, and vehicles in motion. Therefore, our current set of results are generally consistent with our previous study that identified functional subdomains within scene-selective ROIs.

Functional Heterogeneity across Cerebral Cortex

We expect functional heterogeneity to be a prevalent feature across the cerebral cortex, rather than being restricted to the scene selective areas. Therefore, we examined the distribution of networks within FFA, EBA, MT+, and pSTS. In FFA, we find that the human activity (61% ± 10%, $p < 0.05$) and the social interaction (23% ± 6%, $p < 0.05$) (representing scene categories related to verbal communication and entertainment; see Figure 6) are the two leading networks. In EBA and MT+, the human activity network and the motion-energy network occupy a relatively larger portion compared to the remaining networks (44% ± 7% in EBA, $p < 0.05$; 38% ± 6% in MT+, $p < 0.05$). In pSTS, voxels are more broadly distributed across the human activity, social interaction, and civilization networks (22% ± 8%, $p < 0.05$; 36% ± 8%, $p < 0.05$; and 30% ± 9%, $p < 0.05$; respectively). This result is in line with previous findings that pSTS is selective for visual stimuli related to social perception, including the perception of faces, biological motion, others' actions and mental states, and linguistic processing (Deen et al., 2015).

In addition, we examined functional heterogeneity in LO. LO has been associated with the representation of a number of different scene categories (Grill-Spector, Kourtzi, & Kanwisher, 2001; Kim & Biederman, 2011; Lowe, Rajsic, Gallivan, Ferber, & Cant, 2017). In line with previous findings, we find that voxels in LO are broadly distributed across the human activity, motion-energy, texture, and non-human animal networks (37% ± 4%, $p < 0.05$; 30% ± 3%, $p < 0.05$; 16% ± 5%, $p < 0.05$; and 10% ± 4%, $p < 0.05$; respectively).

Next, we examined retinotopically-organized visual areas that are known to be selective for low-level structural features (Grill-Spector & Malach, 2004). As expected, we find that the motion-energy network predominates in V7 (54% ± 4%, $p < 0.05$). However, a large number of voxels in V7 fall into the navigation, non-human animal, and texture networks (16% ± 7%, $p < 0.05$; 13% ± 6%, $p < 0.05$; and 11% ± 4%, $p < 0.05$; respectively). Meanwhile, early visual areas V1–3 are largely dominated by the texture network (68% ± 8%, $p < 0.05$).

In summary, scene-category tuning that we measured within functional ROIs is largely consistent with previously reported response profiles of these areas. Yet, we find that none of the examined ROIs represent a single network. Instead, multiple networks are present in all the ROIs tested. This result implies that selectivity of voxels within conventional ROIs is more diverse than commonly assumed, thus providing further support for the functional heterogeneity hypothesis (Çukur et al., 2016; Çukur, Huth, et al., 2013).

Control Analyses

Theoretical and behavioral accounts suggest that scene categories are partly correlated with global spatial features of natural scenes such as openness, expansion, or roughness, and that human observers might leverage these properties to rapidly categorize visual scenes (Greene & Oliva, 2009; Oliva & Torralba, 2006). Neuroimaging studies have also debated whether PPA represents scene categories or rather correlated low-level spatial features that differ systematically across scene categories (Kravitz et al., 2011; Park et al., 2011; Watson et al., 2014). If the features of the scene-category model are partly correlated with low-level spatial features, then the scene-category model estimated in PPA, OPA, and RSC might be biased.

To rule out this potential confound, we performed several control analyses. First, we fit a separate control model—the gist model—that measures tuning for spatial layout features (Oliva & Torralba, 2001). We find that while the gist model provides significant prediction scores in place-selective ROIs (0.14 ± 0.03 in PPA, 0.15 ± 0.02 in OPA, and 0.13 ± 0.04 in RSC; jackknife test, $p < 0.05$ (FDR corrected)), the scene-category model is superior to the gist model in each ROI (0.38 ± 0.03 in PPA, 0.41 ± 0.01 in OPA, and 0.35 ± 0.03 in RSC; jackknife test, $p < 0.05$). Second, to ensure that heterogeneity of scene-category tuning across the cerebral cortex is not biased by heterogeneity of tuning for low-level spatial features, we compared the average prediction score of the scene-category model and the gist model within the brain networks that were identified by the scene-category model. We find that the scene-category model outperforms the gist model in all networks (jackknife test, $p < 0.05$ (FDR corrected), $\Delta r > 0.02$), except for the texture network that mainly spans across retinotopically-organized early visual areas (jackknife test, $p = 0.23$; Figure 8). This finding is consistent with the notion that early visual areas respond preferentially to low-level spatial features in natural scenes, whereas downstream visual areas respond preferentially to high-level features including scene categories (Grill-Spector & Malach, 2004). Taken together, our results suggest that tuning for low-level spatial features captured by the gist model cannot fully account for scene-category tuning in scene-selective areas, and more broadly across the cerebral cortex.

Previous evidence suggests that scenes that are classified into the same basic-level category (e.g., street, beach, mountain) by human observers tend to possess characteristic

distributions across scene gist features (Oliva & Torralba, 2006). We therefore performed another control analysis to assess the degree of correlation between the scene categories and the gist features. To characterize the distribution of gist features of visual scenes, we performed principal component analysis (PCA) on gist features of natural images from a large public database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). The first 20 gist PCs were retained that explained more than 80% of the total variance of gist features in the movies. The movies were projected onto the gist PCs, and Pearson's correlation was then measured between individual scene categories and gist PCs (Figure 9). On average, we find no significant correlation between scene categories and gist PCs ($|r| = 0.05$; $p = 0.35$, bootstrap test). Only few scene categories (verbal communication, mountain view, pedestrian, text) show modest correlations with the first five gist PCs. To ensure that the scene-category models are not biased by these correlations, we performed a control analysis where we included nuisance regressors that characterized the time course of the first five gist PCs. Scene-category models were fit, and voxel cluster centers were computed. We find that the cluster centers obtained with and without nuisance regressors are nearly identical ($r = 0.93 \pm 0.01$, mean ± sem across subjects; bootstrap test, $p < 0.05$). This result affirms that the differences in scene-category tuning between the brain networks cannot be explained by tuning for low-level spatial features in the stimulus.

Discussion

The aim of this study was to investigate representation of dynamic visual scenes across the human brain. To do this, we fit a scene-category model to measure voxelwise tuning for hundreds of scene categories, where categories were learned inductively as statistical ensembles of objects and actions in natural scenes. We find that this scene-category model explains a significant portion of the response variance broadly across cerebral cortex. We then performed cluster analysis on voxelwise tuning profiles across cortex. Consistently across subjects, we find nine spatially-segregated networks of brain regions that differ in terms of their scene-category tuning. These networks represent a broad variety of natural scene categories related to navigation, human activity, social interaction, civilization, natural environment, non-human animals, motion-energy, and texture.

At a rudimentary level, our results on the spatial organization of scene-category representation suggest a certain degree of functional specialization (Kanwisher, 2010). For instance, areas selective for natural scenes (PPA, OPA, and RSC) are within the navigation network; areas selective for human faces (FFA), bodies (EBA), and social interaction (pSTS) are within the human activity, social interaction, and civilization networks. Yet, an in-depth examination reveals that many conventional ROIs show significant functional heterogeneity. In particular, high-tier areas including FFA, EBA, MT+, RSC, and pSTS each comprise several functional subdomains with differential tuning for scene categories. This result is consistent with two recent studies from our lab analyzing the same natural movies dataset considered here that have identified spatially-segregated subdomains in FFA, PPA, OPA, and RSC, with differential tuning for object categories (Çukur et al., 2016; Çukur, Huth, et al., 2013). Çukur et al. had identified two subdomains within PPA, OPA, and RSC, which differentially responded to dynamic vs. static scenes (Çukur et al., 2016). Similarly, we

observe clusters within RSC, and to a lesser extent in PPA and OPA, that differentially respond to scenes related to navigation (i.e., dynamic) and civilization (i.e., static).

Large parts of not only visual but also nonvisual cortex have been shown to be semantically selective to various categories of objects and actions in a previous study that performed voxelwise modeling on the natural movies dataset analyzed here (Huth et al., 2012). In that previous study, 1705 salient object and action categories in the movies were labeled manually, and then used as binary stimulus features during model fitting. Here we instead used a descriptive learning algorithm to construct scene categories on the basis of detailed co-occurrence statistics of 5252 common objects and actions that were carefully compiled based on movie descriptions provided by Amazon Mechanical Turk workers and a large text corpus. Each scene category was taken as a 5252-dimensional vector containing the probability of occurrence for individual objects and actions within that category. These descriptive features likely increased our sensitivity to capture differences in selectivity for distinct scene categories. For instance, considering a scene where a woman is holding an umbrella while crossing the street on a rainy day and another scene where a woman is holding a cell phone and talking to a man, ‘woman’ and ‘hold’ would come across as two salient features commonly present in both scenes. A model that measures selectivity for salient objects and actions would then predict highly similar responses to these scenes. In contrast, here we can observe a higher-level functional division between these two scenes as the former would elicit responses from navigation and civilization clusters while the latter would elicit responses from human activity and social interaction clusters. This result suggests that scene categories represent nonlinear features beyond a simple linear superposition of objects and actions. While the model proposed by Huth et al. can be more sensitive to changes at the object/action level, our model is more sensitive to changes at the scene-category level.

Another recent study showed that the anterior visual cortex represents scene categories that capture co-occurrence statistics of objects in a large collection of natural images (Stansbury et al., 2013). Here, in addition to using dynamic movies instead of static images, we have also taken into account actions (not only objects) while determining our scene-category features. Furthermore, while Stansbury et al. mostly focused on anterior visual cortex and Çukur et al. focused either on FFA or classical scene-selective regions, here we further considered MT+ and pSTS (Çukur et al., 2016; Çukur, Huth, et al., 2013; Stansbury et al., 2013). This allowed us to identify additional functional subdivisions according to scene category tuning: social interaction, human activity and civilization networks in pSTS and human activity and motion-energy networks in MT+. Finally, we were able to identify a more varied selection of action-related scene categories such as ‘locomotion’, ‘sports activity’, and ‘pedestrian’ whereas Stansbury et al. only had a broad ‘people moving’ category. Hence, our study is based on a fundamentally more diverse set of scene categories that include actions derived from dynamic movies in addition to objects and that are represented not only in anterior visual cortex but also in MT+ and pSTS.

Here we focused on the representation of scene categories based on co-occurrence statistics of objects and actions. However, it has been suggested that at least some scene categories might have discriminating structural features (Oliva & Torralba, 2001, 2006). Therefore, it

is possible that scene-selective areas do not represent scene categories exclusively, but also other structural features that might be systematically related to scene categories (Andrews, Watson, Rice, & Hartley, 2015; Watson et al., 2014). To rule out this confound, we ran control analyses showing that scene category tuning in scene-selective areas cannot be attributed to tuning for low-level spatial texture and layout features, and that heterogeneity of scene-category tuning across neocortex cannot be simply explained by heterogeneity of tuning for these low-level features. Our results add to a growing body of evidence that suggests that high-tier visual areas yield differential responses to images of distinct scene categories, even when the stimuli are controlled to minimize potential correlations between high- and low-level features (Schindler & Bartels, 2016). That said, it is difficult to compile natural stimuli in which feature correlations are completely removed, so it is inherently challenging to disentangle the contributions of low- and high-level features to the organization of scene category representation (Groen, Silson, & Baker, 2017; Lescroart et al., 2015). Further research is required to examine whether, and to what extent, other structural features such as subjective spatial distance (Lescroart et al., 2015), distance to and orientation of large surfaces (Lescroart & Gallant, 2019), spatial expanse (Kravitz et al., 2011; Op de Beeck, Haushofer, & Kanwisher, 2008), or space-defining properties (Mullally & Maguire, 2011) contribute to the representation of scene categories.

In this study, we leveraged co-occurrence statistics of objects and actions to investigate the organization of scene category representation across the cerebral cortex. Because early visual areas predominantly represent low-level visual features of scenes, such as contrast and texture, a scene-category model may not be ideally suited to these areas. A comprehensive assessment of scene representation across the entire cerebral cortex thus requires a hierarchical model that contains features ranging from elementary visual properties to object parts, entire objects, and up to scene categories. Recent studies have utilized convolutional neural networks to extract hierarchical features of natural visual scenes (Agrawal, Stansbury, Malik, & Gallant, 2014; Cadieu et al., 2014; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Groen et al., 2018; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). These studies compared network features at different levels in terms of their success in predicting responses across visual cortex, and they reported that the optimal network level progressively increases towards later visual areas. Although our scene-category model does not explicitly leverage low-level features, the functional organization revealed by voxelwise scene-category tuning profiles is consistent with the cortical hierarchy suggested by these previous reports.

In the scene-category model, each category is defined according to a canonical set of objects and actions that typically occur within that category. Yet, previous reports suggest that not only the category of objects but also the spatial distribution of objects within a scene alters responses in scene-selective areas (Green & Hummel, 2006; Kim & Biederman, 2011; Kim, Biederman, & Juan, 2011). Unlike Stansbury et al., joint consideration of object and action categories in our model carries some implicit information about the spatial distribution of objects (e.g., woman + drive + car versus woman + load + car). However, explicit incorporation of scene layout features would likely further help improve model performance.

A number of cortical networks identified in this study manifest tuning for contemporary scene categories such as “car driven on a road”. It is unlikely for evolution to have sculpted representations of categories that humans have started encountering in present-day environments. Note, however, that visual representations in the human brain do not solely reflect evolution-driven hard-wired aspects of sensory processing. Instead, they also reflect influences from circumstantial factors including sensory experience and functional affordance (Barrett, 2012). Expertise in discriminating exemplars of specific visual categories is thought to alter cortical representations (Tanaka & Curran, 2001). For example, expertise for cars and birds has been associated with increased responses to these objects in FFA (Bilalić, 2016; Gauthier, Skudlarski, Gore, & Anderson, 2000). Furthermore, many brain regions examined here have been implicated in the representation of functional affordances of objects and scenes. For instance, RSC has been reported to represent whether a scene boundary impedes potential navigation (Ferrara & Park, 2016), whereas PPA has been linked to object and scene texture representation (Lowe et al., 2017). Therefore, it is likely that the human brain develops through experience and environmental interactions to code ecologically-important, contemporary scene categories. Yet, we cannot definitively rule out the possibility that scene representations might be influenced by intermediate visual features beyond those examined here. Future studies on scene representation are warranted to shed further light on this issue.

Representation of low-level structural aspects of visual stimuli (e.g., depth and texture) is largely driven by automatic, bottom-up processing (Andrews et al., 2015). In contrast, representation of high-level semantic aspects is influenced by higher cognitive processes and semantic abstraction (Henderson & Hollingworth, 1999). To elicit robust responses from high-level brain regions, here we used an engaging natural movie stimulus, and our subjects were all trained psychophysical observers. Still, some subjects might have inherently maintained lower vigilance than others, which could contribute to across-subject variability. In particular, the scene-category model showed relatively lower performance in S2 compared to remaining subjects (see Table 1). In control analyses, we compared the performance of the scene-category model against the gist model that measures tuning for low-level spatial features. We find that the scene-category model yields substantially higher performance than the gist model in all subjects, except S2 for which the performance improvements with the scene-category model are relatively lower. Thus, a potential explanation for apparent variability in S2 is relatively limited high-level engagement during movie watching.

The VM framework aims to sensitively measure tuning profiles of single voxels in individual subjects. For the natural vision experiment conducted here, the tuning profiles are characterized over a high-dimensional space containing hundreds of scene-category features. To maximize sensitivity of VM models, we conducted prolonged experiments in individual subjects extending over multiple scan sessions. This procedure substantially increases the amount and diversity of fMRI data collected per subject, and enhances the quality of resulting VM models (Çelik et al., 2019). At the same time, given experimental limits, it inevitably constrains the number of subjects that can be recruited. While inclusion of additional subjects might help improve statistical power, the current set of results presented were observed to be highly robust across subjects. We conducted an a priori power analysis

to draw robust inferences about statistical assessment of prediction scores. We found that the outcome of this analysis justifies the results presented in this study.

In summary, we find that cortical areas in visual and nonvisual cortex show heterogeneous tuning for a diverse set of scene categories, and that they are clustered into nine functional networks according to scene-category selectivity. These findings primarily indicate a broader organization of scene representation across the cerebral cortex than typically assumed. Our results also add to a growing body of evidence suggesting a systematic functional organization based on a multi-dimensional semantic space spreading across and extending beyond conventional functional ROIs (Haxby et al., 2011; Huth et al., 2012). The current study supports the idea that information about statistical ensembles of objects and actions is an important contributing factor to the semantic space.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

The authors declare no competing financial interests. The work was supported in part by a National Eye Institute Grant (EY019684), by a Marie Curie Actions Career Integration Grant (PCIG13-GA-2013–618101), by a European Molecular Biology Organization Installation Grant (IG 3028), by a TUBA GEBIP 2015 fellowship, and by a Science Academy BAGEP 2017 award. We thank D. Stansbury, A. Huth, and S. Nishimoto for assistance in various aspects of this research. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedures or analyses was pre-registered prior to the research being conducted.

References

Agrawal P, Stansbury DE, Malik J, & Gallant JL (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. ArXiv.

Aguirre GK, & D'Esposito M (1997). Environmental Knowledge Is Subserved by Separable Dorsal/Ventral Neural Areas. Journal of Neuroscience, 17(7), 2512–2518. 10.1523/JNEUROSCI.17-07-02512.1997 [PubMed: 9065511]

Andrews TJ, Watson DM, Rice GE, & Hartley T (2015). Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. Journal of Vision, 15(7), 3–3. 10.1167/15.7.3

Arthur D, & Vassilvitskii S (2007). k-means++: The Advantages of Careful Seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Barrett HC (2012, 6 26). A hierarchical model of the evolution of human brain specializations. Proceedings of the National Academy of Sciences of the United States of America. 10.1073/pnas.1201898109

Benjamini Y, & Yekutieli D (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics, 29(4), 1165–1188. Retrieved from http://www.jstor.org/stable/2674075

Bilalić M (2016). Revisiting the Role of the Fusiform Face Area in Expertise. Journal of Cognitive Neuroscience, 28(9), 1345–1357. 10.1162/jocn_a_00974 [PubMed: 27082047]

Bird S, Klein E, & Loper E (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.

Blei DM, Ng AY, & Jordan M (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, … DiCarlo JJ (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Computational Biology, 10(12), e1003963. 10.1371/journal.pcbi.1003963 [PubMed: 25521294]

Calvert G, & Campbell R (2003). Reading Speech from Still and Moving Faces: The Neural Substrates of Visible Speech. Journal of Cognitive Neuroscience, 15(1), 57–70. 10.1162/089892903321107828 [PubMed: 12590843]

Cavanna AE, & Trimble MR (2006). The precuneus: A review of its functional anatomy and behavioural correlates. Brain, 129(3), 564–583. 10.1093/brain/awl004 [PubMed: 16399806]

Çelik E, Dar SUH, Yılmaz Ö, Keleş Ü, & Çukur T (2019). Spatially informed voxelwise modeling for naturalistic fMRI experiments. NeuroImage, 186, 741–757. 10.1016/j.neuroimage.2018.11.044 [PubMed: 30502444]

Chen DL, & Dolan WB (2011). Collecting Highly Parallel Data for Paraphrase Evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 190–200). Association for Computational Linguistics. Retrieved from http://www.google.com/transliterate

Cichy RM, Khosla A, Pantazis D, Torralba A, & Oliva A (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific Reports, 6, 27755. 10.1038/srep27755 [PubMed: 27282108]

Çukur T, Huth AG, Nishimoto S, & Gallant JL (2013). Functional Subdomains within Human FFA. Journal of Neuroscience, 33(42), 16748–16766. 10.1523/JNEUROSCI.1259-13.2013 [PubMed: 24133276]

Çukur T, Huth AG, Nishimoto S, & Gallant JL (2016). Functional Subdomains within Scene-Selective Cortex: Parahippocampal Place Area, Retrosplenial Complex, and Occipital Place Area. Journal of Neuroscience, 36(40), 10257–10273. 10.1523/JNEUROSCI.4033-14.2016 [PubMed: 27707964]

Çukur T, Nishimoto S, Huth AG, & Gallant JL (2013). Attention during natural vision warps semantic representation across the human brain. Nature Neuroscience, 16(6), 763. 10.1038/nn.3381 [PubMed: 23603707]

Culham JC, & Kanwisher NG (2001). Neuroimaging of cognitive functions in human parietal cortex. Current Opinion in Neurobiology, 11(2), 157–163. 10.1016/S0959-4388(00)00191-4 [PubMed: 11301234]

David SV, & Gallant JL (2005). Predicting neuronal responses during natural vision. Network: Computation in Neural Systems, 16(2–3), 239–260. 10.1080/09548980500464030

Deen B, Koldewyn K, Kanwisher N, & Saxe R (2015). Functional organization of social perception and cognition in the superior temporal sulcus. Cerebral Cortex, 25(11), 4596–4609. 10.1093/cercor/bhv111 [PubMed: 26048954]

Dilks DD, Julian JB, Paunov AM, & Kanwisher N (2013). The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. Journal of Neuroscience, 33(4), 1331–1336. 10.1523/JNEUROSCI.4081-12.2013 [PubMed: 23345209]

Downing P, Jiang Y, Shuman M, & Kanwisher N (2001). A Cortical Area Selective for Visual Processing of the Human Body. Science, 293(5539), 2470–2473. 10.1126/science.1063414 [PubMed: 11577239]

Eickenberg M, Gramfort A, Varoquaux G, & Thirion B (2017). Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage, 152, 184–194. 10.1016/j.neuroimage.2016.10.001 [PubMed: 27777172]

Engel SA, Glover GH, & Wandell BA (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cerebral Cortex, 7(2), 181–192. 10.1093/cercor/7.2.181 [PubMed: 9087826]

Epstein RA (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. Trends in Cognitive Sciences, 12(10), 388–396. 10.1016/j.tics.2008.07.004 [PubMed: 18760955]

Epstein RA (2014). Neural Systems for Visual Scene Recognition. In Kveraga K & Bar M (Eds.), Scene Vision: Making Sense of What We See (pp. 105–134). Cambridge, MA: MIT Press.

Epstein RA, & Baker CI (2019). Scene Perception in the Human Brain. Annu. Rev. Vis. Sci. 10.1146/annurev-vision-091718

Epstein RA, & Morgan LK (2012). Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis. Neuropsychologia, 50(4), 530–543. 10.1016/j.neuropsychologia.2011.09.042 [PubMed: 22001314]

Epstein R, & Kanwisher N (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598. 10.1038/33402 [PubMed: 9560155]

Ester EF, Sprague TC, & Serences JT (2020). Categorical Biases in Human Occipitoparietal Cortex. The Journal of Neuroscience, 40(4), 917–931. 10.1523/JNEUROSCI.2700-19.2019 [PubMed: 31862856]

Ferrara K, & Park S (2016). Neural representation of scene boundaries. Neuropsychologia, 89, 180–190. 10.1016/j.neuropsychologia.2016.05.012 [PubMed: 27181883]

Gao JS, Huth AG, Lescroart MD, & Gallant JL (2015). Pycortex: an interactive surface visualizer for fMRI. Frontiers in Neuroinformatics, 9, 23. 10.3389/fninf.2015.00023 [PubMed: 26483666]

Gauthier I, Skudlarski P, Gore JC, & Anderson AW (2000). Expertise for cars and birds recruitsbrain areas involved in facerecognition. Nature Neuroscience, 3(2), 191–197. [PubMed: 10649576]

Gauthier I, Tarr MJ, Moylan J, Skudlarski P, Gore JC, & Anderson AW (2000). The fusiform “face area” is part of a network that processes faces at the individual level. Journal of Cognitive Neuroscience, 12(3), 495–504. 10.1162/089892900562165 [PubMed: 10931774]

Green C, & Hummel JE (2006). Familiar interacting object pairs are perceptually grouped. Journal of Experimental Psychology: Human Perception and Performance, 32(5), 1107. 10.1037/0096-1523.32.5.1107 [PubMed: 17002525]

Greene MR, Baldassano C, Esteva A, Beck DM, & Fei-Fei L (2016). Visual Scenes are Categorized by Function. Journal of Experimental Psychology: General, 145(1), 82. 10.1037/xge0000129 [PubMed: 26709590]

Greene MR, & Oliva A (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cognitive Psychology, 58(2), 137–176. 10.1016/j.cogpsych.2008.06.001 [PubMed: 18762289]

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. 10.1016/S0042-6989(01)00073-6 [PubMed: 11322983]

Grill-Spector K, & Malach R (2004). The Human Visual Cortex. Annual Review of Neuroscience, 27, 649–677. 10.1146/annurev.neuro.27.070203.144220

Grill-Spector K, & Weiner KS (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nature Reviews Neuroscience, 15(8), 536. 10.1038/nrn3747 [PubMed: 24962370]

Groen IIA, Greene MR, Baldassano C, Fei-Fei L, Beck DM, & Baker CI (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. ELife, 7, e32962. 10.7554/eLife.32962 [PubMed: 29513219]

Groen IIA, Silson EH, & Baker CI (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1714), 20160102. 10.1098/rstb.2016.0102

Güçlü U, & van Gerven MAJ (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. Journal of Neuroscience, 35(27), 10005–10014. 10.1523/JNEUROSCI.5023-14.2015 [PubMed: 26157000]

Han K, Wen H, Shi J, Lu KH, Zhang Y, Fu D, & Liu Z (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. NeuroImage, 198, 125–136. 10.1016/j.neuroimage.2019.05.039 [PubMed: 31103784]

Hansen KA, Kay KN, & Gallant JL (2007). Topographic Organization in and near Human Visual Area V4. Journal of Neuroscience, 27(44), 11896–11911. 10.1523/JNEUROSCI.2991-07.2007 [PubMed: 17978030]

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, … Ramadge PJ (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron, 72(2), 404–416. 10.1016/j.neuron.2011.08.026 [PubMed: 22017997]

Henderson JM, & Hollingworth A (1999). High-Level Scene Perception. Annual Review of Psychology, 50(1), 243–271. 10.1146/annurev.psych.50.1.243

Huth AG, Nishimoto S, Vu AT, & Gallant JL (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron, 76(6), 1210–1224. 10.1016/j.neuron.2012.10.014 [PubMed: 23259955]

Isik L, Koldewyn K, Beeler D, & Kanwisher N (2017). Perceiving social interactions in the posterior superior temporal sulcus. Proceedings of the National Academy of Sciences, . 10.1073/pnas.1714471114

James G, Witten D, Hastie T, & Tibshirani R (2013). An Introduction to Statistical Learning. New York: Springer.

Jenkinson M, Bannister P, Brady M, & Smith S (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage, 17(2), 825–841. 10.1016/S1053-8119(02)91132-8 [PubMed: 12377157]

Jung Y, Larsen B, & Walther DB (2018). Modality-Independent Coding of Scene Categories in Prefrontal Cortex. Journal of Neuroscience, 38(26), 5969–5981. 10.1523/JNEUROSCI.0272-18.2018 [PubMed: 29858483]

Kable JW, & Chatterjee A (2006). Specificity of Action Representations in the Lateral Occipitotemporal Cortex. Journal of Cognitive Neuroscience, 18(9), 1498–1517. 10.1162/jocn.2006.18.9.1498 [PubMed: 16989551]

Kanwisher N (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. Proceedings of the National Academy of Sciences, 107(25), 11163–11170. 10.1073/pnas.1005062107

Kanwisher N, McDermott J, & Chun MM (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. Journal of Neuroscience, 17(11), 4302–4311. 10.3410/f.717989828.793472998 [PubMed: 9151747]

Khaligh-Razavi SM, & Kriegeskorte N (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Computational Biology, 10(11), e1003915. 10.1371/journal.pcbi.1003915 [PubMed: 25375136]

Kim JG, & Biederman I (2011). Where do objects become scenes? Cerebral Cortex, 21(8), 1738–1746. 10.1093/cercor/bhq240 [PubMed: 21148087]

Kim JG, Biederman I, & Juan C (2011). The Benefit of Object Interactions Arises in the Lateral Occipital Cortex Independent of Attentional Modulation from the Intraparietal Sulcus: A Transcranial Magnetic Stimulation Study. Journal of Neuroscience, 31(22), 8320–8324. 10.1523/JNEUROSCI.6450-10.2011 [PubMed: 21632952]

Konkle T, Brady TF, Alvarez GA, & Oliva A (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. Psychological Science, 21(11), 1551–1556. 10.1177/0956797610385359 [PubMed: 20921574]

Kravitz DJ, Peng CS, & Baker CI (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. 10.1523/JNEUROSCI.4588-10.2011 [PubMed: 21593316]

Lee D, & Seung H (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788. 10.1038/44565 [PubMed: 10548103]

Lescroart MD, & Gallant JL (2019). Human Scene-Selective Areas Represent 3D Configurations of Surfaces. Neuron, 101(1), 178–192. 10.1016/j.neuron.2018.11.004 [PubMed: 30497771]

Lescroart MD, Stansbury DE, & Gallant JL (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Frontiers in Computational Neuroscience, 9, 135. 10.3389/fncom.2015.00135 [PubMed: 26594164]

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, … Zitnick CL (2014). Microsoft COCO: Common Objects in Context. In European conference on computer vision (pp. 740–755). Springer, Cham.

Lowe MX, Rajsic J, Gallivan JP, Ferber S, & Cant JS (2017). Neural representation of geometry and surface properties in object and scene perception. NeuroImage, 157, 586–597. 10.1016/j.neuroimage.2017.06.043 [PubMed: 28647484]

Maguire EA (2001). The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. Scandinavian Journal of Psychology, 42(3), 225–238. 10.1111/1467-9450.00233 [PubMed: 11501737]

Maguire EA, Burgess N, Donnett JG, Frackowiak RS, Frith CD, & O'Keefe J (1998). Knowing Where and Getting There: A Human Navigation Network. Science, 280(5365), 921–924. 10.1126/science.280.5365.921 [PubMed: 9572740]

Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, … Tootell RBH (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. PNAS, 92(18), 8135–8139. 10.1073/pnas.92.18.8135 [PubMed: 7667258]

Malcolm GL, Groen IIA, & Baker CI (2016). Making Sense of Real-World Scenes. Trends in Cognitive Sciences, 20(11), 843–856. 10.1016/j.tics.2016.09.003 [PubMed: 27769727]

Miller G (1995). WordNet: A Lexical Database for English. Communications of the ACM, 38(11), 39–41. 10.1145/219717.219748

Mruczek REB, Von Loga IS, & Kastner S (2013). The representation of tool and non-tool object information in the human intraparietal sulcus. J Neurophysiol, 109, 2883–2896. 10.1152/jn.00658.2012.-Humans [PubMed: 23536716]

Mullally SL, & Maguire EA (2011). A New Role for the Parahippocampal Cortex in Representing Space. Journal of Neuroscience, 31(20), 7441–7449. 10.1523/JNEUROSCI.0267-11.2011 [PubMed: 21593327]

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, & Gallant JL (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Current Biology, 21(19), 1641–1646. 10.1016/j.cub.2011.08.031 [PubMed: 21945275]

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. 10.1023/A:1011139631724

Oliva A, & Torralba A (2006). Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research. 10.1016/S0079-6123(06)55002-2

Op de Beeck HP, Haushofer J, & Kanwisher NG (2008). Interpreting fMRI data: Maps, modules and dimensions. Nature Reviews Neuroscience, 9(2), 123. 10.1038/nrn2314 [PubMed: 18200027]

Park S, Brady TF, Greene MR, & Oliva A (2011). Disentangling Scene Content from Spatial Boundary: Complementary Roles for the Parahippocampal Place Area and Lateral Occipital Complex in Representing Real-World Scenes. Journal of Neuroscience, 31(4), 1333–1340. 10.1523/JNEUROSCI.3885-10.2011 [PubMed: 21273418]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, … Duchesnay É (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from http://scikit-learn.sourceforge.net.

Phan X, & Nguyen C (2007). GibbsLDA++: AC/C++ implementation of latent Dirichlet allocation (LDA).

Posner MI, Sheese BE, Odludaş Y, & Tang Y (2006). Analyzing and shaping human attentional networks. Neural Networks, 19(9), 1422–1429. 10.1016/j.neunet.2006.08.004 [PubMed: 17059879]

Rousselet GA, Joubert OR, & Fabre-Thorpe M (2005). How long to get to the "gist" of real-world natural scenes? Visual Cognition, 12(6), 852–877. 10.1080/13506280444000553

Saxe R (2006). Uniquely human social cognition. Current Opinion in Neurobiology, 16(2), 235–239. 10.1016/j.conb.2006.03.001 [PubMed: 16546372]

Schindler A, & Bartels A (2016). Visual high-level regions respond to high-level stimulus content in the absence of low-level confounds. NeuroImage, 132, 520–525. 10.1016/j.neuroimage.2016.03.011 [PubMed: 26975552]

Serences JT, & Saproo S (2012). Computational advances towards linking BOLD and behavior. Neuropsychologia, 50(4), 435–446. 10.1016/j.neuropsychologia.2011.07.013 [PubMed: 21840553]

Spiridon M, Fischl B, & Kanwisher N (2006). Location and Spatial Profile of Category-Specific Regions in Human Extrastriate Cortex. Human Brain Mapping, 27(1), 77–89. 10.1002/hbm.20169 [PubMed: 15966002]

Stansbury DE, Naselaris T, & Gallant JL (2013). Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. Neuron, 79(5), 1025–1034. 10.1016/j.neuron.2013.06.034 [PubMed: 23932491]

Tanaka JW, & Curran T (2001). A Neural Basis for Expert Object Recognition. Psychological Science, 43–47. [PubMed: 11294227]

Tarhan L, & Konkle T (2020). Sociality and interaction envelope organize visual action representations. Nature Communications, 11(1), 1–11. 10.1038/s41467-020-16846-w

Thompson JC, Clarke M, Stewart T, & Puce A (2005). Configural Processing of Biological Motion in Human Superior Temporal Sulcus. Journal of Neuroscience, 25(39), 9059–9066. 10.1523/JNEUROSCI.2129-05.2005 [PubMed: 16192397]

Tootell RBH, Reppas JB, Kwong KK, Malach R, Born RT, Brady TJ, … Belliveaul JW (1995). Functional Analysis of Human MT and Related Visual Cortical Areas Using Magnetic Resonance Imaging. Journal of Neuroscience, 15(4), 3215–3230. 10.1523/JNEUROSCI.15-04-03215.1995 [PubMed: 7722658]

Tsao DY, Moeller S, & Freiwald WA (2008). Comparing face patch systems in macaques and humans. Proceedings of the National Academy of Sciences, 105(49), 19514–19519. 10.1073/pnas.0809662105

Van Essen D, Drury H, Dickson J, Harwell J, Hanlon D, & Anderson C (2001). An Integrated Software Suite for Surface-based Analyses of Cerebral Cortex. Journal of the American Medical Informatics Association, 8(5), 443–459. 10.1136/jamia.2001.0080443 [PubMed: 11522765]

Walther DB, Caddigan E, Fei-Fei L, & Beck DM (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. Journal of Neuroscience, 29(34), 10573–10581. 10.1523/JNEUROSCI.0559-09.2009 [PubMed: 19710310]

Walther DB, Chai B, Caddigan E, Beck DM, & Fei-Fei L (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. Proceedings of the National Academy of Sciences, 108(23), 9661–9666. 10.1073/pnas.1015666108

Watson DM, Hartley T, & Andrews TJ (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. NeuroImage, 99, 402–410. 10.1016/j.neuroimage.2014.05.045 [PubMed: 24862072]

Weiner KS, Sayres R, Vinberg J, & Grill-Spector K (2010). fMRI-Adaptation and Category Selectivity in Human Ventral Temporal Cortex: Regional Differences Across Time Scales. Journal of Neurophysiology, 103(6), 3349–3365. 10.1152/jn.01108.2009 [PubMed: 20375251]

Wen H, Shi J, Zhang Y, Lu KH, Cao J, & Liu Z (2017). Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. Cerebral Cortex, 28(12), 4136–4160. 10.1093/cercor/bhx268

Xiao J, Hays J, Ehinger KA, Oliva A, & Torralba A (2010). SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In Computer vision and pattern recognition (CVPR), 2010 IEEE conference (pp. 3485–3492). IEEE. Retrieved from http://groups.csail.mit.edu/vision/SUN/.

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, & DiCarlo JJ (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS, 111(23), 8619–8624. 10.1073/pnas.1403112111 [PubMed: 24812127]

Highlights

- Latent Dirichlet Allocation (LDA) is used to uncover scene-category features.
- These features reflect statistical ensembles of not only objects, but also actions.
- Nine spatially-segregated cortical networks with heterogeneous scene-category tunings.
- Scene category representation is more complex than typically assumed.

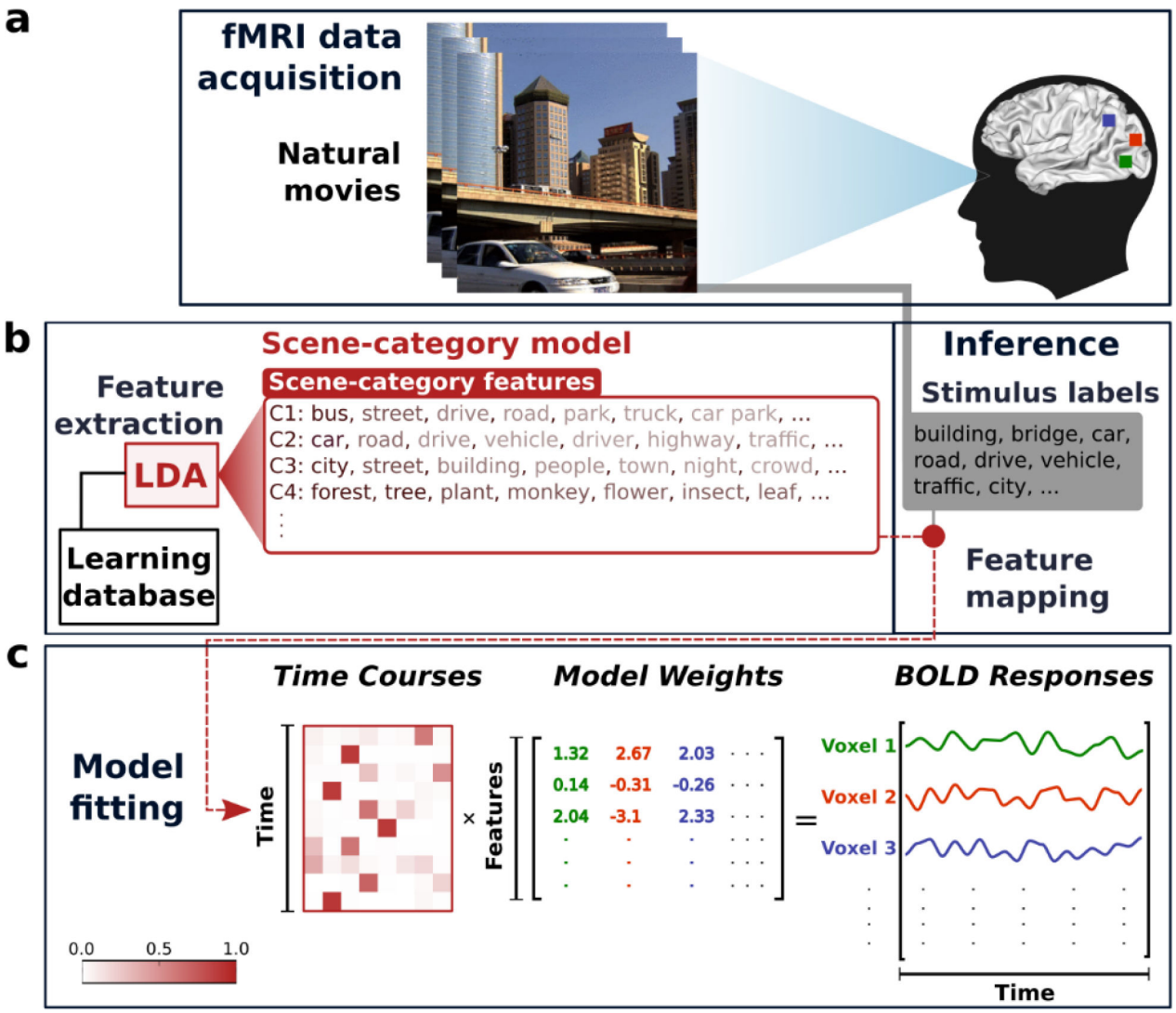

Figure 1. Overview of the voxelwise modeling framework.

a, Whole-brain BOLD responses were recorded while subjects passively viewed two hours of natural movies. **b,** A scene-category model was fit to individual voxels to assess scene category representations of natural movies across neocortex. Model features were extracted via unsupervised learning on a large corpus of natural scene annotations. Scene-category features (C1, C2 etc.) were extracted using latent Dirichlet allocation (LDA) in order to capture co-occurrence statistics of objects and actions in dynamic natural scenes. Each model feature is defined as a list of probabilities that reflect the likelihood of individual objects and actions occurring in a scene. (Font weights for object-action categories reflect their respective probabilities.) **c,** Salient objects and actions in each 1-s clip of the movies were labeled manually. The movies were then projected onto scene-category features to determine stimulus time courses. Regularized linear regression was used to fit voxelwise models that optimally predict BOLD responses in individual voxels. The estimated model weights characterize the tuning of individual voxels for distinct model features.

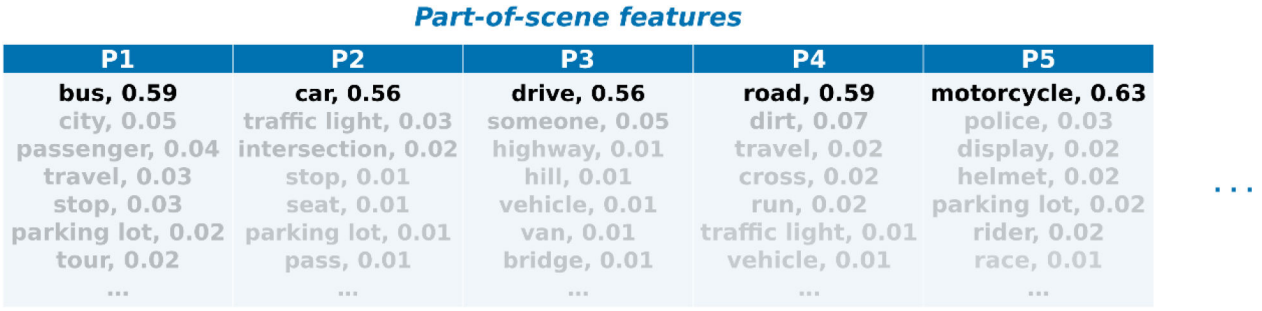

Figure 2. Examples of part-of-scene (P1–P5) features.

The non-negative matrix factorization (NMF) algorithm was used to identify the part-of-scene features that reflect constituent object and action components of natural scenes such as “a bus”, “a car”, “driving”, “road”, and “a motorcycle”. Each model feature was expressed as a 5252-dimensional vector that reflected the probability of occurrence for individual object and action categories. In this example, a total of 80 features were learned. For each feature, seven most probable object and action categories are listed with their probabilities (here indicated as differences in font weights).

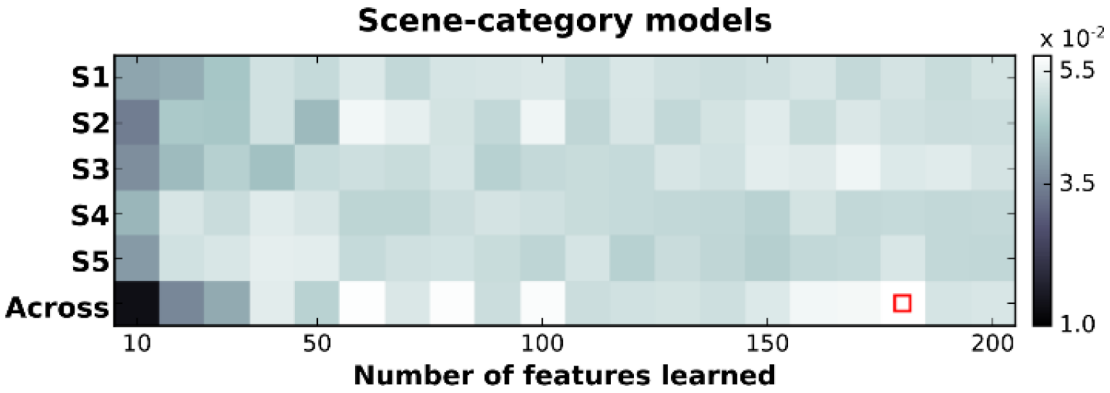

Figure 3. Identifying the set of model features that best explain BOLD responses across subjects.

The number of features to be extracted by LDA was systematically varied from 10 to 200 in 20 steps. For each number of features, voxelwise models were fit and the relative number of cortical voxels that was significantly predicted was measured ($p < 0.05$, FDR corrected). To determine the optimal number of features, these measurements in each subject were normalized to yield a sum of 1 across 20 steps (to account for individual differences in the brain volume and signal-to-noise ratios in BOLD responses). Next, the normalized measurements were averaged across subjects. This matrix shows the number of cortical voxels that were significantly predicted by the scene-category model for individual subjects (S1–S5) and across subjects (Across). The red square indicates the optimal number of features, which is 180 for the scene-category model.

Figure 4. Examples of scene categories that elicit differential responses across voxel clusters.

We determined twenty distinct scene categories frequently observed in daily life that elicited differential responses across voxel clusters. Representative scene categories among this set are shown along with frames from five movie clips with the highest projections onto each category. Scene categories were manually assigned labels to summarize the main scene category information that they captured, including verbal and written communication, entertainment (e.g., playing game, dancing, singing), human female and male, text, anthropomorphic animal, sports activity, locomotion (e.g., jumping, cycling, skiing), pedestrian, car driven on a road, and flying objects (e.g., airplane, insect, bird). These labels were rated by four healthy adult males (non-authors) as a reliability measure.

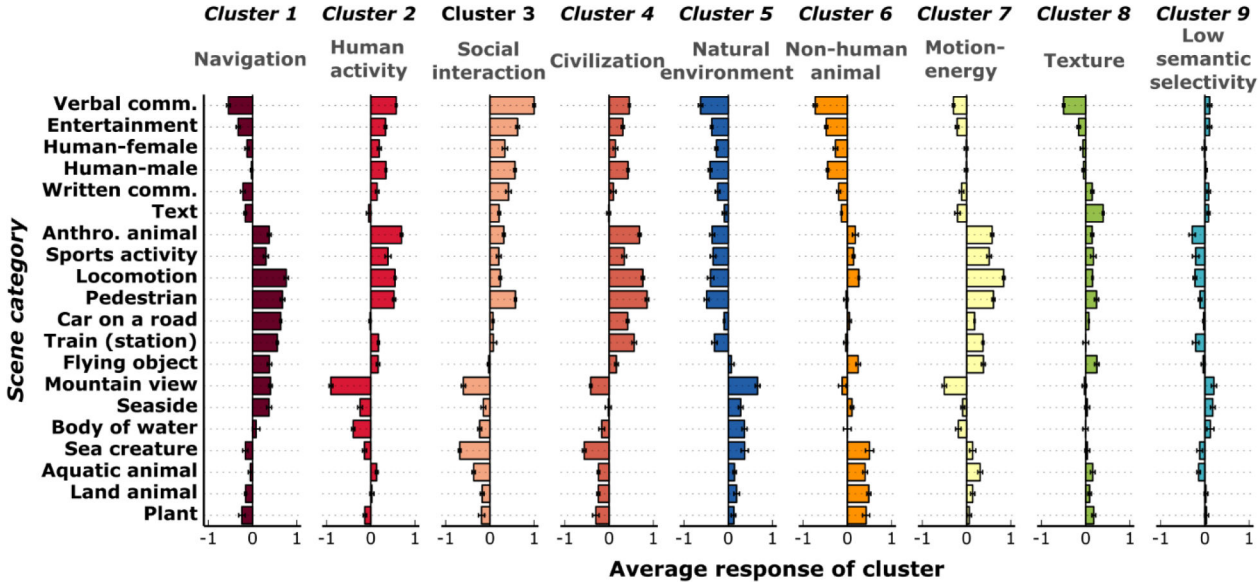

Figure 5. Predicted responses to scene-category features in voxel clusters.

Clustering of voxels according to their scene-category tuning profiles identified nine functional clusters. To identify scene category information represented in each cluster, we measured the average predicted response (mean ± sem across subjects) of each cluster to 60 scene-category features. Results are shown for a subset of scene categories that capture the key response differences across clusters. Cluster 1 (dark brown) responds to scene categories depicting navigation (e.g., human locomotion and vehicles) and landscapes (e.g., a mountain or seaside view). Clusters 2–4 respond to humans and human-made environments: Cluster 2 responds to actions of humans and anthropomorphic animals; Cluster 3 responds to human communication and broadly to social interactions; Cluster 4 responds to human-made environments and artifacts such as vehicles. In contrast, Cluster 5 responds broadly to natural environments, while Cluster 6 responds to non-human animals. Cluster 7 responds to motion-energy in the movies and Cluster 8 responds to texture in visual scenes. Cluster 9 contains the voxels that showed low scene category selectivity in this experiment. The clusters were manually assigned names to reflect their response characteristics.

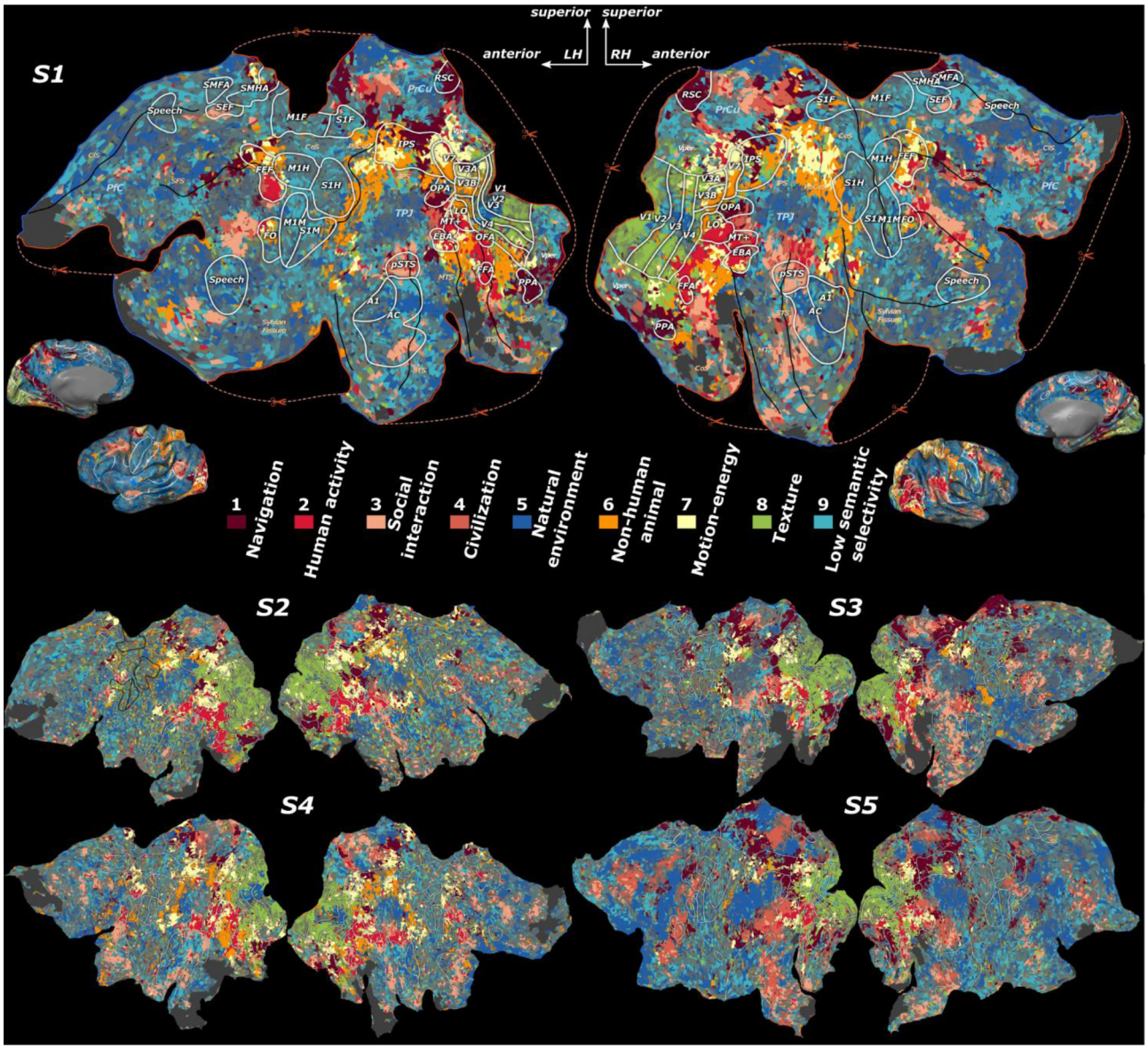

Figure 6. Organization of scene category representation across neocortex.

To examine the cortical distribution of scene category representation, each cortical voxel in each subject was assigned a color according to its cluster membership. Voxel colors were then projected onto cortical flatmaps. Projections obtained for all five subjects are shown. Scene category representation is organized into nine spatially-segregated networks of brain regions. Dark-brown voxels represent scene categories related to navigation (see Figure 5). Red-blush voxels represent scene categories related to human activity, social interaction, and civilization (e.g., human-made artifacts and environments). Dark-blue and orange voxels represent scene categories related to natural environments and non-human animals. Light-blue voxels show low scene category selectivity. Voxels that were not significantly predicted by the scene-category model are shown in gray.

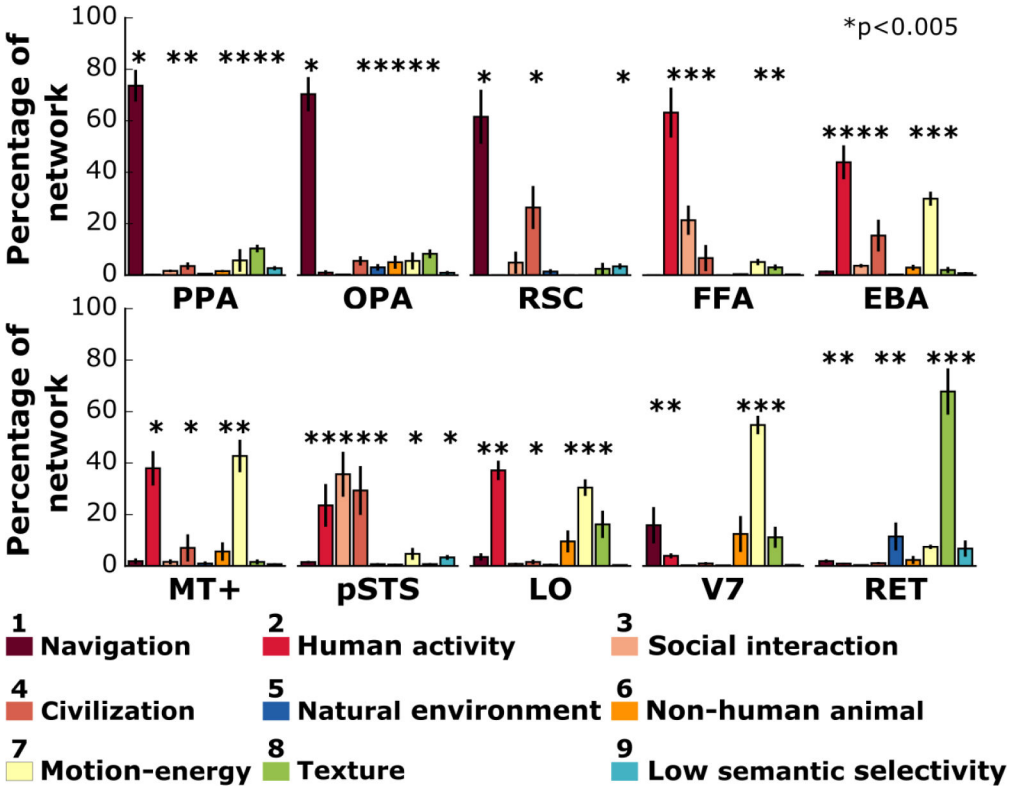

Figure 7. Distribution of networks of brain regions within conventional ROIs.

Previous work from our lab (Çukur et al., 2016; Çukur, Huth, et al., 2013) suggests that voxels within ROIs defined by conventional localizers do not belong to a single network, but rather are associated with multiple different networks. To address this question, we assessed the distribution of networks within PPA, OPA, RSC, FFA, EBA, MT+, pSTS, LO, V7, and RET. Bar plots for each ROI indicate the percentages of voxels (mean ± sem across subjects) that belong to the nine networks. Asterisks indicate percentages that are significantly different than zero ($p < 0.05$, bootstrap test, FDR corrected). Most of the ROIs examined here contain multiple subdomains with distinct scene-category tuning. This result suggests that scene category representations in many functional ROIs are more diverse than commonly assumed.

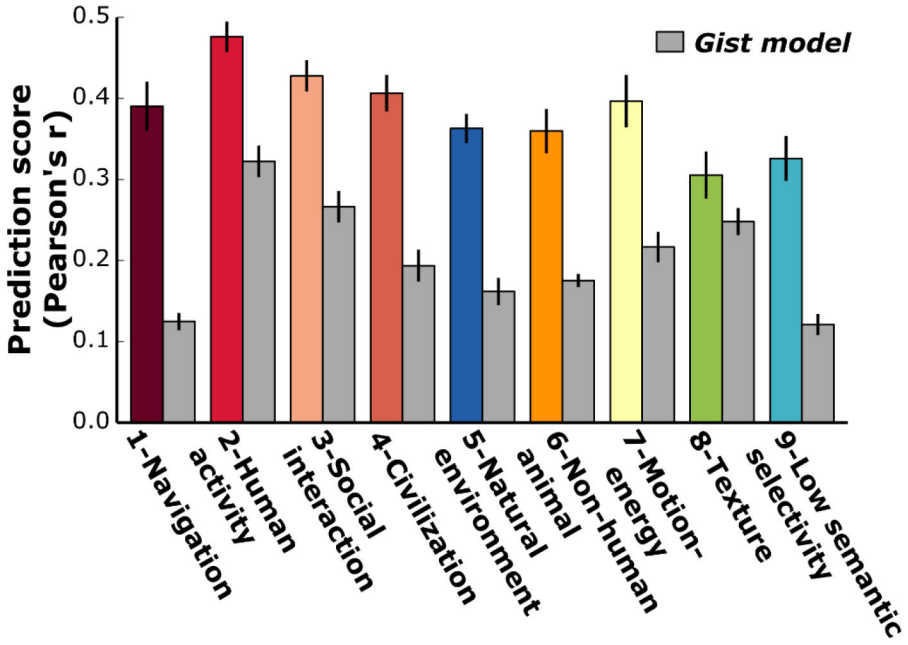

Figure 8. The prediction scores of scene-category and gist models in each network.

To control whether heterogeneity of scene-category tuning across neocortex was biased by heterogeneity of tuning for low-level spatial features of natural scenes, we compared the prediction scores of the scene-category and gist models within the networks that were identified by the scene-category model. Bar plots show prediction scores (mean ± sem across subjects). The scene-category model outperforms the gist model in all networks (jackknife test, $p < 0.05$, $\Delta r > 0.02$), except for the texture network that mainly spans across early visual areas (jackknife test, $p = 0.23$). Note that the largest difference in prediction scores is observed within the navigation network, which was found to largely overlap with scene-selective areas PPA, OPA, and RSC. This result suggests that differences in scene-category tuning between the identified networks cannot be fully attributed to tuning for low-level spatial features.

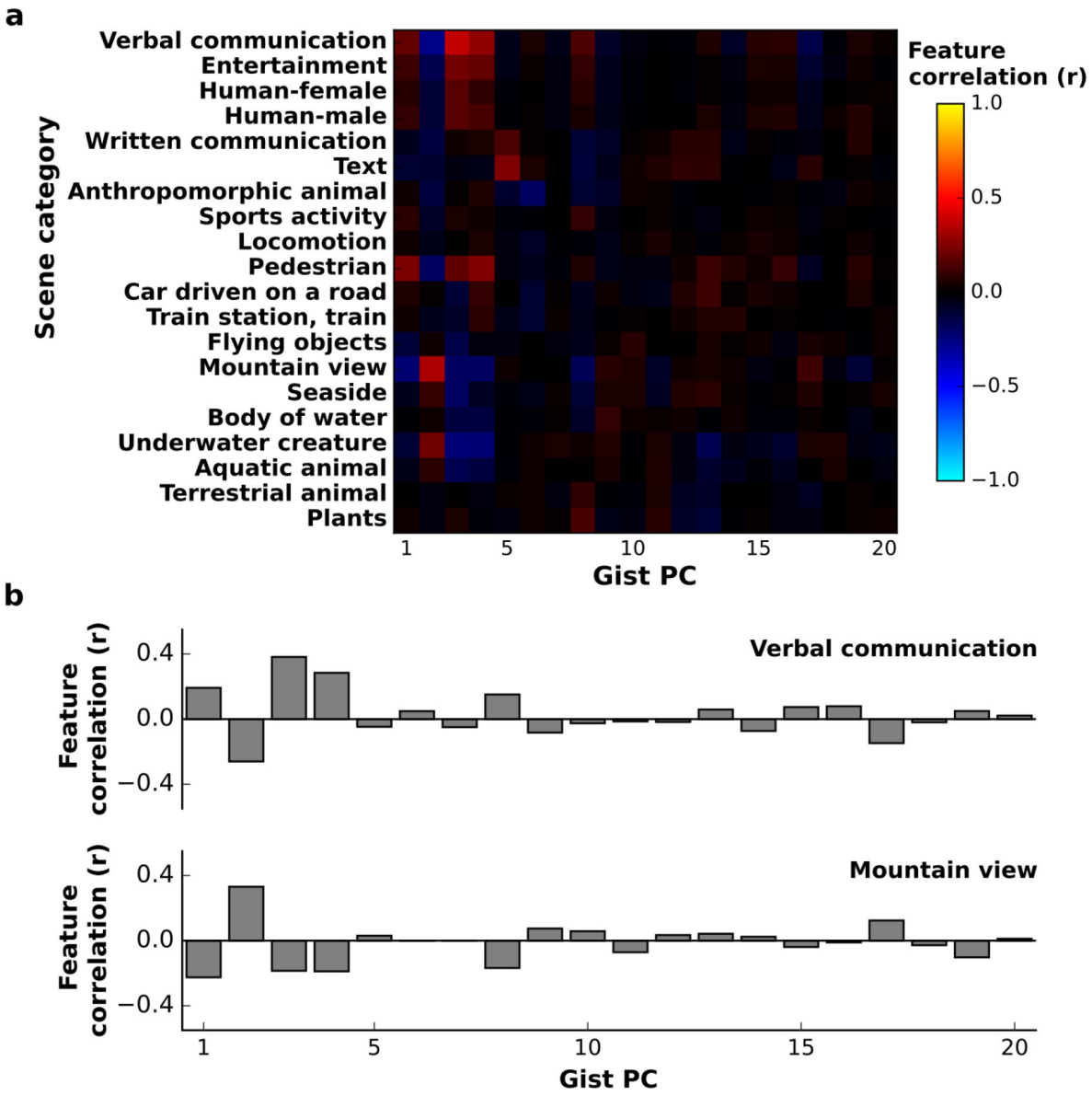

Figure 9. Correlation between the scene-category features and the gist features.

Heterogeneity of scene-category tuning across neocortex could potentially be biased by heterogeneity of tuning for low-level spatial features that differ systematically across scene-categories. To rule out this bias, the degree of correlation between the 20 scene-category features and the first 20 PCs of gist features was calculated. **a,** Correlation among scene categories and gist PCs presented in matrix form (see color legend). A few scene categories including verbal communication, mountain view, pedestrian, and text show moderate correlations with the first five gist PCs (see bottom panel for a bar plot of correlations of the verbal communication and mountain view categories with the gist PCs). These PCs assess global spatial properties of natural scenes, such as roughness, openness, verticalness, mean depth, and expansion. Even so, the majority of scene categories that elicit differential responses between the networks have negligible correlations with the gist PCs.

Table 1.

Proportion of voxels where the scene-category model outperforms the control model.

% voxels	S1	S2	S3	S4	S5	Aggregate
RET	53.68	27.26	24.76	33.43	74.44	42.71±9.42
V7	56.96	5.71	16.48	63.35	39.06	36.31±11.17
LO	50.00	5.26	3.45	87.93	65.00	42.33±16.64
MT+	49.42	8.54	22.92	56.36	48.97	37.24±9.17
EBA	53.38	2.40	11.84	64.81	45.06	35.50±12.09
FFA	16.13	7.32	6.25	52.17	30.77	22.53±8.61
PPA	88.46	62.71	72.92	90.00	74.56	77.73±5.12
OPA	79.41	6.00	46.15	100.00	90.74	64.46±17.22
RSC	65.45	91.30	82.43	79.17	68.57	77.38±4.70
IPS	61.33	18.53	40.43	76.67	69.91	53.37±10.63
pSTS	25.34	24.13	33.45	63.83	47.52	38.85±7.51

Percentage of voxels where the scene-category model outperforms the control model among voxels significantly predicted by either of the two models is calculated for eleven ROIs for all subjects ($\Delta r > 0.02$). Aggregate values are reported as mean ± SEM across five subjects.