HHS Public Access

Author manuscript

J Exp Psychol Gen. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

J Exp Psychol Gen. 2016 January ; 145(1): 82–94. doi:10.1037/xge0000129.

Visual Scenes are Categorized by Function

Michelle R. Greene[1],[2], **Christopher Baldassano**[1], **Andre Esteva**[3], **Diane M. Beck**[4], and **Li Fei-Fei**[1]

[1]Stanford University, Department of Computer Science

[2]Minerva Schools at KGI

[3]Stanford University, Department of Electrical Engineering

[4]University of Illinois at Urbana-Champaign

Abstract

How do we know that a kitchen is a kitchen by looking? Traditional models posit that scene categorization is achieved through recognizing necessary and sufficient features and objects, yet there is little consensus about what these may be. However, scene categories should reflect how we use visual information. We therefore test the hypothesis that scene categories reflect functions, or the possibilities for actions within a scene. Our approach is to compare human categorization patterns with predictions made by both functions and alternative models. We collected a large-scale scene category distance matrix (5 million trials) by asking observers to simply decide whether two images were from the same or different categories. Using the actions from the American Time Use Survey, we mapped actions onto each scene (1.4 million trials). We found a strong relationship between ranked category distance and functional distance (r=0.50, or 66% of the maximum possible correlation). The function model outperformed alternative models of object-based distance (r=0.33), visual features from a convolutional neural network (r=0.39), lexical distance (r=0.27), and models of visual features. Using hierarchical linear regression, we found that functions captured 85.5% of overall explained variance, with nearly half of the explained variance captured only by functions, implying that the predictive power of alternative models was due to their shared variance with the function-based model. These results challenge the dominant school of thought that visual features and objects are sufficient for scene categorization, suggesting instead that a scene's category may be determined by the scene's function.

Keywords

scene understanding; categorization; similarity

Author for correspondence: Michelle R. Greene, Department of Computer Science, Room 240, Stanford University, 353 Serra Mall, Stanford, CA 94305.

Introduction

> "The question 'What makes things seem alike or different?' is one so fundamental to psychology that very few psychologists have been naïve enough to ask it"
>
> (Attneave, 1950).

Although more than half a century has passed since Attneave issued this challenge, we still have little understanding of how we categorize and conceptualize visual content. The notion of similarity, or family resemblance, is implicit in how content is conceptualized (Wittgenstein, 2010), yet similarity cannot be defined except in reference to a feature space to be operated over (Goodman, 1972; Medin, Goldstone, & Gentner, 1993). What feature spaces determine environmental categories? Traditionally, it has been assumed that this feature space is comprised of a scene's component visual features and objects (Biederman, 1987; Bulthoff, Edelman, & Tarr, 1995; Marr, 1982; Riesenhuber & Poggio, 1999; Stansbury, Naselaris, & Gallant, 2013). Mounting behavioral evidence, however, indicates that human observers have high sensitivity to the global meaning of an image (Fei-Fei, Iyer, Koch, & Perona, 2007; Greene & Oliva, 2009a, 2009b; Potter, 1976), and very little sensitivity to the local objects and features that are outside the focus of attention (Rensink, 2002). Consider the image of the kitchen in Figure 1. If objects determine scene category membership, then we would expect the kitchen supply store (left) to be conceptually equivalent to the kitchen. Alternatively, if scenes are categorized (labeled) according to spatial layout and surfaces (Bar, 2004; Oliva & Torralba, 2001; Torralba, Fergus, & Freeman, 2008), then observers might place the laundry room (center) into the same category as the kitchen. However, most of us share the intuition that the medieval kitchen (right) is in the same category, despite sharing few objects and features with the top image. Why is the image on the right a better category match to the modern kitchen than the other two?

Here we put forth the hypothesis that the conceptual structure of environments is driven primarily by the scene's *functions*, or the actions that one could perform in the scene. We assert that representing a scene in terms of its high-level functions is a better predictor of how humans categorize scenes than state-of-the-art models representing a scene's visual features or objects.

Figure 2 illustrates our approach. We constructed a large-scale scene category distance matrix by querying over 2,000 observers on over 63,000 images from 1055 scene categories (Figure 2A). Here, the distance between two scene categories was proportional to the number of observers who indicated that the two putative categories were "different". We compared this human categorization pattern with an function-based pattern created by asking hundreds of observers to indicate which of several hundred actions could take place in each scene (Figure 2B). We can then compute the function-based distance for each pair of categories. We found a striking resemblance between function-based distance and the category distance pattern. The function model not only explained more variance in the category distance matrix than leading models of visual features and objects, but also contributed the most uniquely explained variance of any tested model. These results suggest that a scene's functions provide a fundamental coding scheme for human scene

categorization. In other words, of the models tested, the functions afforded by the scene best explains why we consider two images to be from the same category.

Methods

Creating Human Scene Category Distance Matrix

The English language has terms for hundreds of types of environments, a fact reflected in the richness of large-scale image databases such as ImageNet (Jia Deng et al., 2009) or SUN (Xiao, Ehinger, Hays, Torralba, & Oliva, 2014). These databases used the WordNet (Miller, 1995) hierarchy to identify potential scene categories. Yet we do not know how many of these categories reflect basic- or entry-level scene categories, as little is known about the hierarchical category structure of scenes (Tversky & Hemenway, 1983). Therefore, our aim was to discover this category structure for human observers at a large scale.

To derive a comprehensive list of scene categories, we began with a literature review. Using Google Scholar, we identified 116 papers in human visual cognition, cognitive neuroscience, or computer vision matching the keywords “scene categorization” or “scene classification” that had a published list of scene categories. 1535 unique category terms were identified over all papers. Our goal was to identify scene categories with at least 20 images in publically available databases. We removed 204 categories that did not meet this criterion. We then removed categories describing animate entities (e.g. “Crowd of people”, N=44); specific places (e.g. “Alaska”, N=42); events (e.g. “forest fire”, N=35); or objects (e.g. “playing cards”, N=93). Finally, we omitted 62 categories for being close synonyms of another (e.g. “country” and “countryside”). This left us with a total of 1055 scene categories. To obtain images for each category, 722 categories were found in the SUN database (Xiao et al., 2014), 306 were taken from ImageNet (Jia Deng et al., 2009), 24 from the Corel database, and three from the 15-scene database of (Fei-Fei & Perona, 2005; Lazebnik, Schmid, & Ponce, 2006; Oliva & Torralba, 2001).

We will refer to the 1,055 scene categories as putative categories. Good categories have both high within-category similarity (cohesion), as well as high between-category distance (distinctiveness) (Iordan, Greene, Beck, & Fei-Fei, 2015; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). We performed a large-scale experiment with over 2,000 human observers using Amazon’s Mechanical Turk (AMT). In each trial, two images were presented to observers side by side. Half of the image pairs came from the same putative scene category, while the other half were from two different categories that were randomly selected. Image exemplars were randomly selected within a category on each trial. In order to encourage participants to categorize at the basic- or entry-level (Jolicoeur, Gluck, & Kosslyn, 1984; Tversky & Hemenway, 1983), we gave participants the following instructions: “Consider the two pictures below, and the names of the places they depict. Names should describe the type of place, rather than a specific place and should make sense in finishing the following sentence ’I am going to the ….’ ”, following the operational definition applied in the creation of the SUN database (Xiao et al., 2014). To ensure that the instructions were understood and followed, participants were also asked to type in the category name that they would use for the image on the left-hand side. These data were not

analyzed. Participants were not placed under time pressure to respond, and images remained on screen until response was recorded.

Potential participants were recruited from a pool of trusted observers with at least 2,000 previously approved trials with at least 98% approval. Additionally, participants were required to pass a brief scene vocabulary test before participating. In the vocabulary test, potential participants were required to match ten scene images to their appropriate category name (see Supplementary Material for names and images). 245 potential participants attempted the qualification test and did not pass. Trials from 14 participants were omitted from analysis for inappropriate typing in the response box. Trials were omitted when workers pasted the image URL into the category box instead of providing a name (N=586 trials from 3 workers), for submitting the hit before all trials were complete (N=559 trials from 4 workers), for typing category names in languages other than English (N=195 trials from 2 workers), typing random character strings (N=111 trials from 2 workers), or for typing in words such as “same”, “left”, or “pictures”, implying that the instructions were not understood (N=41 trials from 3 workers). Workers were compensated $0.02 for each trial. We obtained at least 10 independent observations for each cell in the 1055 by 1055 scene matrix, for a total of over 5 million trials. Individual participants completed a median of 5 hits of this task (range: 1–36,497). There was a median of 1,116 trials in each of the diagonal entries of the matrix, and a median of 11 trials in each cell of the off-diagonal entries.

From the distribution of “same” and “different” responses, we created a dissimilarity matrix in which the distance between two scene categories was defined as the proportion of participants who indicated that the two categories were “different”. From the 1,055 categories, we identified 311 categories with the strongest within-category cohesion (at least 70% of observers agreed that images were from the same category). In general, categories that were omitted were visually heterogeneous, such as “community center”, or were inherently multimodal. For example, “dressing room” could reflect the backstage area of a theatre, or a place to try on clothes in a department store. Thus, the final dataset included 311 scene categories from 885,968 total trials, and from 2,296 individual workers.

Creating the Scene Function Spaces

In order to determine whether scene categories are governed by functions, we needed a broad space of possible actions that could take place in our comprehensive set of scene categories. We gathered these actions from the lexicon of the American Time Use Survey (ATUS), a project sponsored by the US Bureau of Labor Statistics that uses U.S. census data to determine how people distribute their time across a number of activities. The lexicon used in this study was pilot tested over the course of three years (Shelley, 2005), and therefore represents a complete set of goal-directed actions that people can engage in. This lexicon was created independently from any question surrounding vision, scenes, or categories, therefore avoiding the potential problem of having functions that were designed to distinguish among categories of visual scenes. Instead, they simply describe common actions one can engage in in everyday life. The ATUS lexicon includes 428 specific activities organized into 17 major activity categories and 105 mid-level categories. The 227

actions included in our study included the most specific category levels with the following exceptions:

1. The superordinate category “Caring for and Helping Non-household members” was dropped as these actions would be visually identical to those in the “Caring for and Helping Household members” category.
2. In the ATUS lexicon, the superordinate-level category “Work” contained only two specific categories (primary and secondary jobs). Because different types of work can look very visually different, we expanded this category by adding 22 categories representing the major labor sectors from the Bureau of Labor Statistics.
3. The superordinate-level category “Telephone calls” was collapsed into one action because we reasoned that all telephone calls would look visually similar.
4. The superordinate-level category “Traveling” was similarly collapsed into one category because being in transit to go to school (for example) should be visually indistinguishable from being in transit to go to the doctor.
5. All instances of “Security procedures” have been unified under one category for similar reasons.
6. All instances of “Waiting” have been unified under one category.
7. All “Not otherwise specified” categories have been removed.

The final list of actions can be found in the Supplemental Materials.

To compare this set of comprehensive functions to a human-generated list of functions applied to visual scenes, we took the 36 function/affordance rankings from the SUN attribute database (Patterson, Xu, Su, & Hays, 2014). In this set, observers were asked to generate attributes that differentiated scenes.

Mapping Functions Onto Images

In order to test our hypothesis that scene category distance is reflected in the distance of scenes’ functions, we need to map functions onto scene categories. Using a separate large-scale online experiment, 484 participants indicated which of the 227 actions could take place in each of the 311 scene categories. Participants were screened using the same criterion described above. In each trial, a participant saw a randomly selected exemplar image of one scene category along with a random selection of 17 or 18 of the 227 actions. Each action was hyperlinked to its description in the ATUS lexicon. Participants were instructed to use check boxes to indicate which of the actions would typically be done in the type of scene shown.

Each individual participant performed a median of 9 trials (range: 1–4,868). Each scene category – function pair was rated by a median of 16 participants (range: 4–86), for a total of 1.4 million trials.

We created a 311-category by 227-function matrix in which each cell represents the proportion of participants indicating that the action could take place in the scene category.

Since scene categories varied widely in the number of actions they afford, we created a distance matrix by computing the cosine distance between all possible pairs of categories, resulting in a 311×311 function-based distance matrix. This measures the overlap between actions while being invariant to the absolute magnitude of the action vector.

Function Space MDS Analysis

To better understand the scene function space, we performed a classical metric multidimensional scaling (MDS) decomposition of the function distance matrix. This yielded an embedding of the scene categories such that inner products in this embedding space approximate the (double-centered) distances between scene categories, with the embedding dimensions ranked in order of importance (Buja et al., 2008). In order to better understand the MDS dimensions, we computed the correlation coefficient between each action (across scene categories) with the category coordinates for a given dimension. This provides us with the functions that are the most and least associated with each dimension.

Alternative Models

To put the performance of the function-based model in perspective, we compared it to nine alternative models based on previously proposed scene category primitives. Five of the models represented visual features, one model considered human-generated scene attributes, and one model examined the human-labeled objects in the scenes. As with the function model, these models yielded scene category by feature matrices that were converted to distance matrices using cosine distance, and then compared to the category distance matrix. The object and attribute models, like the functional model, were created from human observers' scene labeling. Additionally, two models measured distances directly, based either on the lexical distance between scene category names (the Semantic Model), or simply by whether scenes belonged to the same superordinate level category (indoor, urban or natural; the Superordinate-Category Model). We will detail each of the models below.

Models of Visual Features—A common framework for visual categorization and classification involves finding the necessary and sufficient visual features to perform categorization e.g. (Fei-Fei & Perona, 2005; Lazebnik et al., 2006; Oliva & Torralba, 2001; Renninger & Malik, 2004; Vogel & Schiele, 2007). Here we constructed distance matrices based on various visual feature models to determine how well they map on the human categorization (i.e. the category dissimilarity matrix) and in particular compare their performance to our functional category model.

Convolutional Neural Network—In order to represent the state-of-the-art in terms of visual features, we generated a visual feature vector using the publicly distributed OverFeat convolutional neural network (CNN) (Sermanet et al., 2013), which was trained on the ImageNet 2012 training set (Jia Deng et al., 2009). These features, computed by iteratively applying learned nonlinear filters to the image, have been shown to be a powerful image representation for a wide variety of visual tasks (Razavian, Azizpour, Sullivan, & Carlsson, 2014). This 7-layer CNN takes an image of size 231×231 as input, and produces a vector of 4096 image features that are optimized for 1000-way object classification. This network achieves top-5 object recognition on ImageNet 2012 with approximately 16% error,

meaning that the correct object is one of the model's first five responses in 84% of trials. Using the top layer of features, we averaged the features for all images in each scene category to create a 311-category by 4096-feature matrix.

Gist—We used the Gist descriptor features of (Oliva & Torralba, 2001). This popular model for scene recognition provides a summary statistic representation of the dominant orientations and spatial frequencies at multiple scales coarsely localized on the image plane. We used spatial bins at 4 cycles per image and 8 orientations at each of 4 spatial scales for a total of 3,072 filter outputs per image. We averaged the gist descriptors for each image in each of the 311 categories to come up with a single 3,072-dimensional descriptor per category.

Color histograms—In order to determine the role of color similarity in scene categorization, we represented color using LAB color space. For each image, we created a two-dimensional histogram of the a* and b* channels using 50 bins per channel. We then averaged these histograms over each exemplar in each category, such that each category was represented as a 2500 length vector representing the averaged colors for images in that category. The number of bins was chosen to be similar to those used in previous scene perception literature (Oliva & Schyns, 2000).

Tiny Images—Torralba and colleagues (Torralba et al., 2008) demonstrated that human scene perception is robust to aggressive image downsampling, and that an image descriptor representing pixel values from such downsampled images could yield good results in scene classification. Here, we downsampled each image to 32 by 32 pixels (grayscale). We created our 311-category by 1024 feature matrix by averaging the downsampled exemplars of each category together.

Gabor Wavelet Pyramid—To assess a biologically inspired model of early visual processing, we represented each image in this database as the output of a bank of multi-scale Gabor filters. This type of representation has been used to successfully model the representation in early visual areas (Kay, Naselaris, Prenger, & Gallant, 2008). Each image was converted to grayscale, down sampled to 128 by 128 pixels, and represented with a bank of Gabor filters at three spatial scales (3, 6 and 11 cycles per image with a luminance-only wavelet that covers the entire image), four orientations (0, 45, 90 and 135 degrees) and two quadrature phases (0 and 90 degrees). An isotropic Gaussian mask was used for each wavelet, with its size relative to spatial frequency such that each wavelet has a spatial frequency bandwidth of 1 octave and an orientation bandwidth of 41 degrees. Wavelets were truncated to lie within the borders of the image. Thus, each image is represented by 3*3*2*4+6*6*2*4+11*11*2*4 = 1328 total Gabor wavelets. We created the feature matrix by averaging the Gabor weights over each exemplar in each category.

Object-based Model—Our understanding of high-level visual processing has generally focused on object recognition, with scenes considered as a structured set of objects (Biederman, 1987). Therefore, we also consider a model of scene categorization that is explicitly built upon objects. In order to model the similarity of objects within scene categories, we employed the LabelMe tool (Russell et al, 2008) that allows users to outline

and annotate each object in each image by hand. 7,710 scenes from our categories were already labeled in the SUN 2012 release (Xiao et al., 2014), and we augmented this set by labeling an additional 223 images. There were a total of 3,563 unique objects in this set. Our feature matrix consisted of the proportion of scene images in each category containing a particular object. For example, if 10 out of 100 *kitchen* scenes contained a "blender", the entry for kitchen-blender would be 0.10. In order to estimate how many labeled images we would need to robustly represent a scene category, we performed a bootstrap analysis in which we resampled the images in each category with replacement (giving the same number of images per category as in the original analysis), and then measured the variance in distance between categories. With the addition of our extra images, we ensured that all image categories either had at least 10 fully labeled images or had mean standard deviation in distance to all other categories of less than 0.05 (e.g. less than 5% of the maximal distance value of 1).

Scene-Attribute Model—Scene categories from the SUN database can be accurately classified according to human-generated attributes that describe a scene's material, surface, spatial, and functional scene properties (Patterson et al., 2014). In order to compare our function-based model to another model of human-generated attributes, we used the 66 non-function attributes from (Patterson et al., 2014) for the 297 categories that were common to our studies. To further test the role of functions, we then created a separate model from the 36 function-based attributes from their study. These attributes are listed in the Supplementary Material.

Semantic Models—Although models of visual categorization tend to focus on the necessary features and objects, it has long been known that most concepts cannot be adequately expressed in such terms (Wittgenstein, 2010). As semantic similarity has been suggested as a means of solving category induction (Landauer & Dumais, 1997), we examined the extent to which category structure follows from the semantic similarity between category names. We examined semantic similarity by examining the shortest path between category names in the WordNet tree using the Wordnet::Similarity implementation of (Pedersen, Patwardhan, & Michelizzi, 2004). The similarity matrix was normalized and converted into distance. We examined each of the metrics of semantic relatedness implemented in Wordnet::Similarity and found that this path measure was the best correlated with human performance.

Superordinate-Category Model—As a baseline model, we examined how well a model that groups scenes only according to superordinate-level category would predict human scene category assessment. We assigned each of the 311 scene categories to one of three groups (natural outdoors, urban outdoors or indoor scenes). These three groups have been generally accepted as mutually exclusive and unambiguous superordinate-level categories (Tversky & Hemenway, 1983; Xiao et al., 2014). Then, each pair of scene categories in the same group was given a distance of 0 while pairs of categories in different groups were given a distance of 1.

Model Assessment—To assess how each of the feature spaces resembles the human categorization pattern, we created a 311×311 distance matrix representing the distance between each pair of scene categories for each feature space. We then correlated the off-diagonal entries in this distance matrix with those of the category distance matrix from the scene categorization experiment. Since these matrices are symmetric, the off-diagonals were represented in a vector of 48,205 distances.

Noise Ceiling—The variability of human categorization responses puts a limit on the maximum correlation expected by any of the tested models. In order to get an estimate of this maximum correlation, we used a bootstrap analysis in which we sampled with replacement observations from our scene categorization dataset to create two new datasets of the same size as our original dataset. We then correlated these two datasets to one another, and repeated this process 1000 times.

Hierarchical Regression Analysis—In order to understand the unique variance contributed by each of our feature spaces, we used hierarchical linear regression analysis, using each of the feature spaces both alone and in combination to predict the human categorization pattern. In total, 15 regression models were used: (1) all feature spaces used together; (2) the top four performing features together (functions, objects, attributes and the CNN visual features); (3–6) each of the top four features alone; (6–11) each pair of the top four features; (12–15) each set of three of the top four models. By comparing the r^2 values of a feature space used alone to the r^2 values of that space in conjunction with another feature space, we can infer the amount of variance that is independently explained by that feature space. In order to visualize this information in an Euler diagram, we used EulerAPE software (Micallef & Rodgers, 2014).

Results

Human Scene Category Distance

To assess the conceptual structure of scene environments, we asked over 2,000 human observers to categorize images as belonging to 311 scene categories in a large-scale online experiment. The resulting 311 by 311 category distance matrix is shown in Figure 3. In order to better visualize the category structure, we have ordered the scenes using the optimal leaf ordering for hierarchical clustering (Bar-Joseph, Gifford, & Jaakkola, 2001); allowing us to see what data-driven clusters emerge.

Several category clusters are visible. Some clusters appear to group several subordinate-level categories into a single entry-level concept, such as “bamboo forest”, “woodland” and “rainforest” being examples of *forests.* Other clusters seem to reflect broad classes of activities (such as “sports”) which are visually heterogeneous and cross other previously defined scene boundaries, such as indoor-outdoor (Fei-Fei et al., 2007; Henderson, Larson, & Zhu, 2007; Szummer & Picard, 1998; Tversky & Hemenway, 1983), or the size of the space (Greene & Oliva, 2009a; Oliva & Torralba, 2001; Park, Konkle, & Oliva, 2014). Such activity-oriented clusters hint that the actions that one can perform in a scene (the scene’s functions) could provide a fundamental grouping principle for scene category structure.

Function-based Distance Best Correlates with Human Category Distance

For each of our feature spaces, we created a distance vector (see Model Assessment) representing the distance between each pair of scene categories. We then correlated this distance vector with the human distance vector from the previously described experiment.

In order to quantify the performance of each of our models, we defined a noise ceiling based on the inter-observer reliability in the human scene distance matrix. This provides an estimate of the explainable variance in the scene categorization data, and thus provides an upper bound on the performance of any of our models. Using bootstrap sampling (see Methods), we found an inter-observer correlation of r=0.76. In other words, we cannot expect a correlation with any model to exceed this value.

Function-based similarity had the highest resemblance to the human similarity pattern (r=0.50 for comprehensive set, and r=0.51 for the 36 functional attributes). This represents about 2/3 of the maximum observable correlation obtained from the noise ceiling. As shown in Figure 4A, this correlation is substantially higher than any of the alternative models we tested. The two function spaces were highly correlated with one another (r=0.63). As they largely make the same predictions, we will use the results from the 227-function set for the remainder of the paper.

Of course, being able to perform similar actions often means manipulating similar objects, and scenes with similar objects are likely to share visual features. Therefore, we compared function-based categorization patterns to alternative models based on perceptual features, non-function attributes, object-based similarity, and the lexical similarity of category names.

We tested five different models based on purely visual features. The most sophisticated used the top-level features of a state-of-the-art convolutional neural network model (CNN, (Sermanet et al., 2013) trained on the ImageNet database (Jia Deng et al., 2009). Category distances in CNN space produced a correlation with human category dissimilarity of r=0.39. Simpler visual features, however, such as gist (Oliva & Torralba, 2001), color histograms (Oliva & Schyns, 2000), Tiny Images (Torralba et al., 2008), and wavelets (Kay et al., 2008) had low correlations with human scene category dissimilarity.

Category structure could also be predicted to some extent based on the similarity between the objects present in scene images (r=0.33, using human-labeled objects from the LabelMe database, (Russell et al., 2008), the non function-based attributes (r=0.28) of the SUN attribute database (Patterson et al., 2014), or the lexical distance between category names in the WordNet tree (Huth, Nishimoto, Vu, & Gallant, 2012; Miller, 1995; Pedersen et al., 2004)(r=0.27). Surprisingly, a model that merely groups scenes by superordinate-level categories (indoor, urban or natural environments) also had a sizeable correlation (r=0.25) with human dissimilarity patterns.

Although each of these feature spaces had differing dimensionalities, this pattern of results also holds if the number of dimensions is equalized through principal components analysis. We created minimal feature matrices by using the first N PCA components, and then correlated the cosine distance in these minimal feature spaces with the human scene

distances, see Figure 5. We found that the functional features were still the most correlated with human behavior.

Independent Contributions from Alternative Models

To what extent does function-based similarity *uniquely* explain the patterns of human scene categorization? Although function-based similarity was the best explanation of the human categorization pattern of all the models we tested, CNN and object-based models also had sizeable correlations with human behavior. To what extent do these models make the same predictions?

In order to assess the independent contributions made by each of the models, we used a hierarchical linear regression analysis in which each of the three top-performing models was used either separately or in combination to predict the human similarity pattern. By comparing the r^2 values from the individual models to the r^2 values for the combined model, we can assess the unique variance explained by each descriptor. A combined model with all features explained 31% of the variance in the human similarity pattern (r=0.56). This model is driven almost entirely by the top four feature spaces (functions, CNN, attribute, and object labels), which explained 95% of the variance from all features, a combined 29.4% of the total variance (r=0.54). Note that functions explained 85.6% of this explained variance, indicating that the object and perceptual features only added a small amount of independent information (14.4% of the combined variance). Variance explained by all 15 regression models is listed in Table 1.

Although there was a sizable overlap between the portions of the variance explained by each of the models (see Figure 4B), around half of the total variance explained can be attributed *only* to functions (44.2% of the explained variance in top four models), and was not shared by the other three models. In contrast, the independent variance explained by CNN features, object-based features, and attributes accounted for only 6.8%, 0.6%, and 0.4% of the explained variance respectively. Therefore, the contributions of visual, attribute, and object-based features are largely shared with function-based features, further highlighting the utility of functions for explaining human scene categorization patterns.

Functions Explain All Types of Scene Categories

Does the impressive performance of the functional model hold over all types of scene categories, or is performance driven by outstanding performance on a particular type of scene? To address this question, we examined the predictions made by the three top-performing models (functions, CNN and objects) on each of the superordinate-level scene categories (indoor, urban and natural landscape) separately. As shown in Table 2, we found that the function-based model correlated similarly with human categorization in all types of scenes. This is in stark contrast to the CNN and object models, whose performance was driven by performance on the natural landscape scenes.

Examining Scene Function Space

In order to better understand the function space, we performed classical multi-dimensional scaling on the function distance matrix, allowing us to identify how patterns of functions

contribute to the overall similarity pattern. We found that at least 10 MDS dimensions were necessary to explain 95% of the variance in the function distance matrix, suggesting that the efficacy of the function-based model was driven by a number of distinct function dimensions, rather than just a few useful functions. We examined the projection of categories onto the first three MDS dimensions. As shown in Figure 6, the first dimension appears to separate indoor locations that have a high potential for social interactions (such as “socializing” and “attending meetings for personal interest”) from outdoor spaces that afford more solitary activities, such as “hiking” and “science work”. The second dimension separates work-related activities from leisure. Later dimensions appear to separate environments related to transportation and industrial workspaces from restaurants, farming, and other food-related environments, see Figure 7 for listing of associated categories and functions for each MDS dimension. A follow-up experiment demonstrated that functions that are highly associated with a particular object (e.g. “mailing” is strongly associated with objects such as mailboxes and envelopes) are equally predictive of categorization patterns as functions that do not have strong object associates (e.g. “helping an adult”), see Supplementary Materials for details.

Why does the function space have higher fidelity for predicting human patterns of scene categorization? To concretize this result, we will examine a few failure cases for alternative features. Category names should reflect cognitively relevant categories, so what hurts the performance of the lexical distance model? This model considers the categories “access road” and “road tunnel” to have the lowest distance of all category pairs (possibly because both contain the term “road”), while only 10% of human observers placed these into the same category. By contrast, the function model considered them to be rather distant, with only 35% overlap between functions (intersection over union). Shared functions included “in transit / travelling” and “architecture and engineering work”, while tunnels independently afforded “rock climbing and caving” and access roads often contained buildings, thus affording “building grounds and maintenance work”. If objects such as buildings can influence both functions and categories, then why don’t objects fare better? Consider the categories “underwater kelp forest” and “underwater swimming pool”. The object model considers them to be very similar given the presence of water, but 80% of human observers consider them to be different. Similarly, these categories share only 17% overlap in functions, with the kelp forest affording actions such as “science work”, while the swimming pool affords “playing sports with children”.

Of course, certain failure cases of the function model should also be mentioned. For example, while all human observers agreed that “bar” and “tea room” were different categories, the function model considered them to be similar, given their shared functions of “socializing”, “eating and drinking”, “food preparation and serving work” etc. Similarly, the function model considered “basketball arena” and “theatre” to be similar, while human observers did not. Last, the function model also frequently confused scene categories that shared a particular sport, such as “baseball field” and “indoor batting cage”, while no human observers placed them in the same category. However, it should be noted that human observers also shared this last trait in other examples, with 55% of observers placing “bullpen” and “pitcher’s mound” into the same category.

Discussion

We have shown that human scene categorization is better explained by the action possibilities, or functions, of a scene than by the scene's visual features or objects. Furthermore, function-based features explained far more independent variance than did alternative models, as these models were correlated with human category patterns only insofar as they were also correlated with the scene's functions. This suggests that a scene's functions contain essential information for categorization that is not captured by the scene's objects or visual features.

The current results cannot be explained by the smaller dimensionality of the function-based features, as further analysis revealed that function-based features outperformed other feature spaces using equivalent numbers of dimensions. Furthermore, this pattern was observed over a wide range of dimensions, suggesting that each functional feature contained more information about scene categories than each visual or object-based feature. Critically, the function-based model performed with similar fidelity on all types of scenes, which is a hallmark of human scene perception (Kadar & Ben-Shahar, 2012) that is not often captured in computational models. Indeed, indoor scene recognition is often much harder for computer models than other classification problems (Quattoni & Torralba, 2009; Szummer & Picard, 1998) and this was true for our visual and object-based models, while the function model showed high fidelity for explaining indoor scene categorization.

The idea that the function of vision is for action has permeated the literature of visual perception, but it has been difficult to fully operationalize this idea for testing. Psychologists have long theorized that rapid and accurate environmental perception could be achieved by the explicit coding of an environment's affordances, most notably in J.J. Gibson's influential theory of ecological perception (Gibson, 1986). This work is most often associated with the direct perception of affordances that reflect relatively simple motor patterns such as sitting or throwing. As the functions used in the current work often reflect higher-level, goal-directed actions, and because we are making no specific claims about the direct perception of these functions, we have opted not to use the term affordances here. Nonetheless, ideas from Gibson's ecological perception theory have inspired this work, and thus we consider our functions as conceptual extensions of Gibson's idea.

In our work, a scene's functions are those actions that one can imagine doing in the scene, rather than the activities that one reports as occurring in the scene. This distinguishes this work from that of activity recognition (Aggarwal & Ryoo, 2011; Hafri, Papafragou, & Trueswell, 2013; Wiggett & Downing, 2010; Yao & Fei-Fei, 2010), placing it closer to the ideas of Gibson and the school of ecological psychology.

Previous small-scale studies have found that environmental functions such as navigability are reflected in patterns of human categorization (Greene & Oliva, 2009a, 2010), and are perceived very rapidly from images (Greene & Oliva, 2009b). Our current results provide the first comprehensive, data-driven test of this hypothesis, using data from hundreds of scene categories and affordances. By leveraging the power of crowdsourcing, we were able to obtain both a large-scale similarity structure for visual scenes, but also normative ratings

of functions for these scenes. Using hundreds of categories, thousands of observers and millions of observations, crowdsourcing allowed a scale of research previously unattainable. Previous research on scene function has also suffered from the lack of a comprehensive list of functions, relying instead on the free responses of human observers describing the actions that could be taken in scenes (Greene & Oliva, 2009a; Patterson & Hays, 2012). By using an already comprehensive set of actions from the American Time Use Survey, we were able to see the full power of functions for predicting human categorization patterns. The current results speak only to categorization patterns obtained from unlimited viewing times, and future work will examine the extent to which function-based categorization holds for limited viewing times, similar to previous work (Greene & Oliva, 2009a, 2009b).

Given the relatively large proportion of variance independently explained by function-based features, we are left with the question of why this model outperforms the more classic models. By examining patterns of variance in the function by category matrix, we found that functions can be used to separate scenes along previously defined dimensions of scene variance, such as superordinate-level category (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky & Larson, 2010; Tversky & Hemenway, 1983), and between work and leisure activities (Ehinger, Torralba, & Oliva, 2010). Although the variance explained by function-based similarity does not come directly from visual features or the scene's objects, human observers must be able to apprehend these functions from the image somehow. It is therefore a question open for future work to understand the extent to which human observers bring non-visual knowledge to bear on this problem. Of course, it is possible that functions can be used in conjunction with other features for categorization, just as shape can be determined independently from shading (Ramachandran, 1988), motion (Julesz, 1971) or texture (Gibson, 1950).

Some recent work has examined large-scale neural selectivity based on semantic similarity (Huth et al., 2012), or object-based similarity (Stansbury et al., 2013), finding that both types of conceptual structures can be found in the large-scale organization of human cortex. Our current work indeed shows sizeable correlations between these types of similarity structures and human behavioral similarity. However, we find that function-based similarity is a better predictor of behavior and may provide an even stronger grouping principle in the brain.

Despite the impressive predictive power of functions for explaining human scene categorization, many open questions are still left about the nature of functions. To what extent are they perceptual primitives as suggested by Gibson, and to what extent are they inherited from other diagnostic information? The substantial overlap between functions and objects and visual features (Figure 4B) implies that at least some functions are correlated with these features. Intuitively this makes sense as some functions, such as "mailing" may be strongly associated with objects such as a mailbox or an envelope. However, our results suggest that the mere presence of an associated object may not be enough: just because the kitchen supply store has pots and pans does not mean that one can cook there. The objects must conform in type, number, and spatial layout to jointly give rise to functions. Furthermore, some functions such as "jury duty", "waiting", and "socializing" are harder to associate with particular objects and features, and may require higher-level, non-visual

knowledge. While the current results bypass the issue of how observers compute the functions, we must also examine how the functions can be understood directly from images in a bottom-up manner.

These results challenge many existing models of visual categorization that consider categories to be purely a function of shared visual features or objects. Just as the Aristotelian theory of concepts assumed that categories could be defined in terms of necessary and sufficient features, classical models of visual categorization have assumed that a scene category can be explained by necessary and sufficient objects (Biederman, 1987; Stansbury et al., 2013) or diagnostic visual features (Renninger & Malik, 2004; Vogel & Schiele, 2007). However, just as the classical theory of concepts cannot account for important cognitive phenomena, the classical theory of scene categories cannot account for the fact that two scenes can share a category even when they do not share many features or objects. By contrast, the current results demonstrate that the possibility for action creates categories of environmental scenes. In other words, a kitchen is a kitchen because it is a space that affords cooking, not because it shares objects or other visual features with other kitchens.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Aggarwal JK, Ryoo MS. Human Activity Analysis: A Review. ACM Comput Surv. 2011; 43(3):16:1–16:43. http://doi.org/10.1145/1922649.1922653.

Attneave F. Dimensions of similarity. The American Journal of Psychology. 1950; 63(4):516–556. [PubMed: 14790020]

Bar-Joseph Z, Gifford DK, Jaakkola TS. Fast optimal leaf ordering for hierarchical clustering. Bioinformatics. 2001; 17:S22–S29. http://doi.org/10.1093/bioinformatics/17.suppl_1.S22. [PubMed: 11472989]

Bar M. Visual objects in context. Nature Reviews Neuroscience. 2004; 5:617–625. [PubMed: 15263892]

Biederman I. Recognition-by-Components: A Theory of Human Image Understanding. Psychological Review. 1987; 94(2):115–47. [PubMed: 3575582]

Buja A, Swayne DF, Littman ML, Dean N, Hofmann H, Chen L. Data Visualization With Multidimensional Scaling. Journal of Computational and Graphical Statistics. 2008; 17(2):444–472. http://doi.org/10.1198/106186008X318440.

Bulthoff HH, Edelman SY, Tarr MJ. How Are Three-Dimensional Objects Represented in the Brain? Cereb Cortex. 1995; 5(3):247–260. http://doi.org/10.1093/cercor/5.3.247. [PubMed: 7613080]

Ehinger KA, Torralba A, Oliva A. A taxonomy of visual scenes: Typicality ratings and hierarchical classification. Journal of Vision. 2010; 10(7):1237–1237. http://doi.org/10.1167/10.7.1237.

Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? Journal of Vision. 2007; 7(1:10):1–29. [PubMed: 17997664]

Fei-Fei, L.; Perona, P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02; IEEE Computer Society; 2005. p. 524-531.Retrieved from http://portal.acm.org/citation.cfm?id=1069129

Gibson, JJ. The perception of the visual world. Vol. xii. Oxford, England: Houghton Mifflin; 1950.

Gibson, JJ. The Ecological Approach to Visual Perception. Lawrence Erlbaum Associates; 1986.

Goodman, N. Problems and Projects. Bobs-Merril; 1972. Seven Strictures on Similarity.

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

Greene MR, Oliva A. The Briefest of Glances: The Time Course of Natural Scene Understanding. Psychological Science. 2009b; 20:464–472. http://doi.org/10.1111/j.1467-9280.2009.02316.x. [PubMed: 19399976]

Greene MR, Oliva A. High-level aftereffects to global scene properties. Journal of Experimental Psychology: Human Perception and Performance. 2010; 36(6):1430–1432. [PubMed: 20731502]

Hafri A, Papafragou A, Trueswell JC. Getting the gist of events: Recognition of two-participant actions from brief displays. Journal of Experimental Psychology: General. 2013; 142(3):880–905. http://doi.org/10.1037/a0030045. [PubMed: 22984951]

Henderson J, Larson C, Zhu D. Cortical activation to indoor versus outdoor scenes: an fMRI study. Experimental Brain Research. 2007; 179(1):75–84. http://doi.org/10.1007/s00221-006-0766-2. [PubMed: 17123070]

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

Iordan MC, Greene MR, Beck DM, Fei-Fei L. Basic Level Category Structure Emerges Gradually across Human Ventral Visual Cortex. Journal of Cognitive Neuroscience. 2015; 27(7):1427–1446. http://doi.org/10.1162/jocn_a_00790. [PubMed: 25811711]

Deng, Jia; Dong, Wei; Socher, R.; Li, Li-Jia; Li, Kai; Fei-Fei, Li. ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009; IEEE; 2009. p. 248-255.http://doi.org/10.1109/CVPR.2009.5206848

Jolicoeur P, Gluck MA, Kosslyn SM. Pictures and names: making the connection. Cognitive Psychology. 1984; 16(2):243–75. [PubMed: 6734136]

Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M. Processing scene context: Fast categorization and object interference. Vision Research. 2007; 47(26):3286–3297. http://doi.org/10.1016/j.visres.2007.09.013. [PubMed: 17967472]

Julesz, B. Foundations of Cyclopean Perception. MIT Press; 2006.

Kadar I, Ben-Shahar O. A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. Journal of Vision. 2012; 12(13) http://doi.org/10.1167/12.13.16.

Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452(7185):352–355. http://doi.org/10.1038/nature06713. [PubMed: 18322462]

Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review. 1997; 104(2):211–240. http://doi.org/10.1037/0033-295X.104.2.211.

Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2; IEEE Computer Society; 2006. p. 2169-2178.Retrieved from http://portal.acm.org/citation.cfm?id=1153171.1153549

Loschky LC, Larson AM. The natural/man-made distinction is made before basic-level distinctions in scene gist processing. Visual Cognition. 2010; 18(4):513–536. http://doi.org/10.1080/13506280902937606.

Marr, D. Vision. W.H. Freeman; 1982.

Medin DL, Goldstone RL, Gentner D. Respects for similarity. PSYCHOLOGICAL REVIEW. 1993; 100:254–278. http://doi.org/10.1.1.11.6647.

Micallef L, Rodgers P. eulerAPE: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses. PLoS ONE. 2014; 9(7):e101717. http://doi.org/10.1371/journal.pone.0101717. [PubMed: 25032825]

Miller GA. WordNet: a lexical database for English. Commun ACM. 1995; 38(11):39–41. http://doi.org/10.1145/219717.219748.

Oliva A, Schyns PG. Diagnostic colors mediate scene recognition. Cognitive Psychology. 2000; 41:176–210. [PubMed: 10968925]

Oliva A, Torralba A. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision. 2001; 42(3):145–175. http://doi.org/10.1023/A:1011139631724.

Park, S.; Konkle, T.; Oliva, A. Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain; Cerebral Cortex. 2014. p. bht418http://doi.org/10.1093/cercor/bht418

Patterson, G.; Hays, J. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR); 2012.

Patterson G, Xu C, Su H, Hays J. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. International Journal of Computer Vision. 2014; 108(1–2):59–81. http://doi.org/10.1007/s11263-013-0695-z.

Pedersen, T.; Patwardhan, S.; Michelizzi, J. Demonstration Papers at HLT-NAACL 2004. Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. WordNet::Similarity: Measuring the Relatedness of Concepts; p. 38-41.Retrieved from http://dl.acm.org/citation.cfm?id=1614025.1614037

Potter MC. Short-term conceptual memory for pictures. Journal of Experimental Psychology. Human Learning and Memory. 1976; 2(5):509–522. [PubMed: 1003124]

Pylyshyn Z. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. The Behavioral and Brain Sciences. 1999; 22(3):341–365. discussion 366–423. [PubMed: 11301517]

Quattoni, A.; Torralba, A. Recognizing indoor scenes. 2014 IEEE Conference on Computer Vision and Pattern Recognition; Los Alamitos, CA, USA: IEEE Computer Society; 2009. p. 413-420.http://doi.org/10.1109/CVPRW.2009.5206537

Ramachandran VS. Perception of shape from shading. Nature. 1988; 331(6152):163–166. http://doi.org/10.1038/331163a0. [PubMed: 3340162]

Razavian, AS.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features off-the-shelf: an Astounding Baseline for Recognition. 2014. arXiv:1403.6382 [cs]. Retrieved from http://arxiv.org/abs/1403.6382

Renninger LW, Malik J. When is scene indentification just texture recognition? Vision Research. 2004; 44:2301–2311. [PubMed: 15208015]

Rensink RA. Change detection. Annual Review of Psychology. 2002; 53:245–277.

Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nature Neuroscience. 1999; 2(11):1019–1025. http://doi.org/10.1038/14819. [PubMed: 10526343]

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic objects in natural categories. Cognitive Psychology. 1976; 8(3):382–439. http://doi.org/10.1016/0010-0285(76)90013-X.

Russell B, Torralba A, Murphy K, Freeman W. LabelMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision. 2008; 77(1):157–173. http://doi.org/10.1007/s11263-007-0090-8.

Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. 2013. arXiv:1312.6229 [cs]. Retrieved from http://arxiv.org/abs/1312.6229

Shelley KJ. Developing the American Time Use Survey Activity Classification System. Monthly Labor Review. 2005; 128(6):3–15.

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

Szummer, M.; Picard, RW. Indoor-Outdoor Image Classification. Content-Based Access of Image and Video Databases, Workshop on; Los Alamitos, CA, USA: IEEE Computer Society; 1998. p. 42http://doi.org/http://doi.ieeecomputersociety.org/10.1109/CAIVD.1998.64-6032

Torralba A, Fergus R, Freeman WT. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008; 30(11):1958–1970. [PubMed: 18787244]

Tversky B, Hemenway K. Categories of environmental scenes. Cognitive Psychology. 1983; 15:121–149.

Vogel J, Schiele B. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. Int J Comput Vision. 2007; 72(2):133–157.

Wiggett AJ, Downing PE. Representation of Action in Occipito-temporal Cortex. Journal of Cognitive Neuroscience. 2010; 23(7):1765–1780. http://doi.org/10.1162/jocn.2010.21552. [PubMed: 20807060]

Wittgenstein, L. Philosophical Investigations. John Wiley & Sons; 2010.

Xiao, J.; Ehinger, KA.; Hays, J.; Torralba, A.; Oliva, A. SUN Database: Exploring a Large Collection of Scene Categories; International Journal of Computer Vision. 2014. p. 1-20.http://doi.org/10.1007/s11263-014-0748-y

Yao, B.; Fei-Fei, L. Modeling mutual context of object and human pose in human-object interaction activities. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2010. p. 17-24.http://doi.org/10.1109/CVPR.2010.5540235

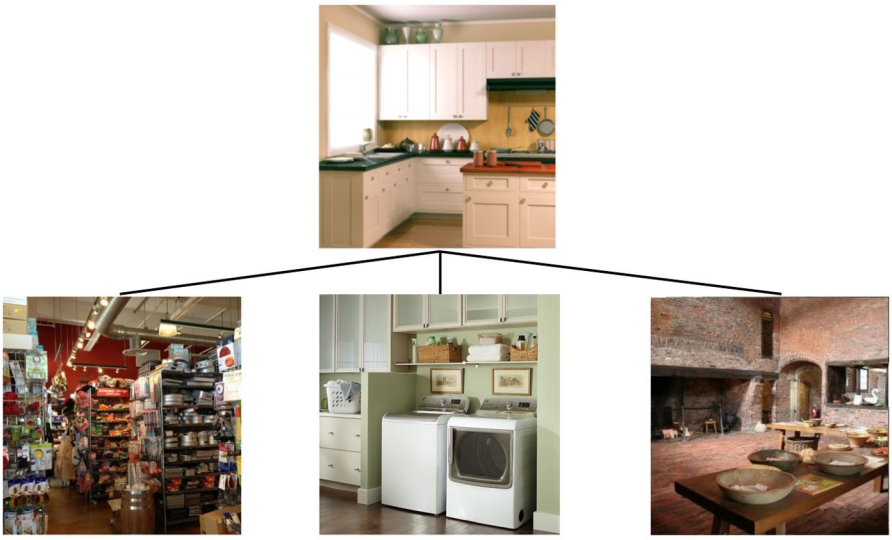

Figure 1.
The top image depicts a kitchen. Which of the bottom images is also a kitchen? Many influential models of visual categorization assume that scenes sharing objects, such as the kitchen supply store (left), or layout, such as the laundry room (middle) would be placed into the same category by human observers. Why is the medieval kitchen also a kitchen despite having very different objects and features from the top kitchen?

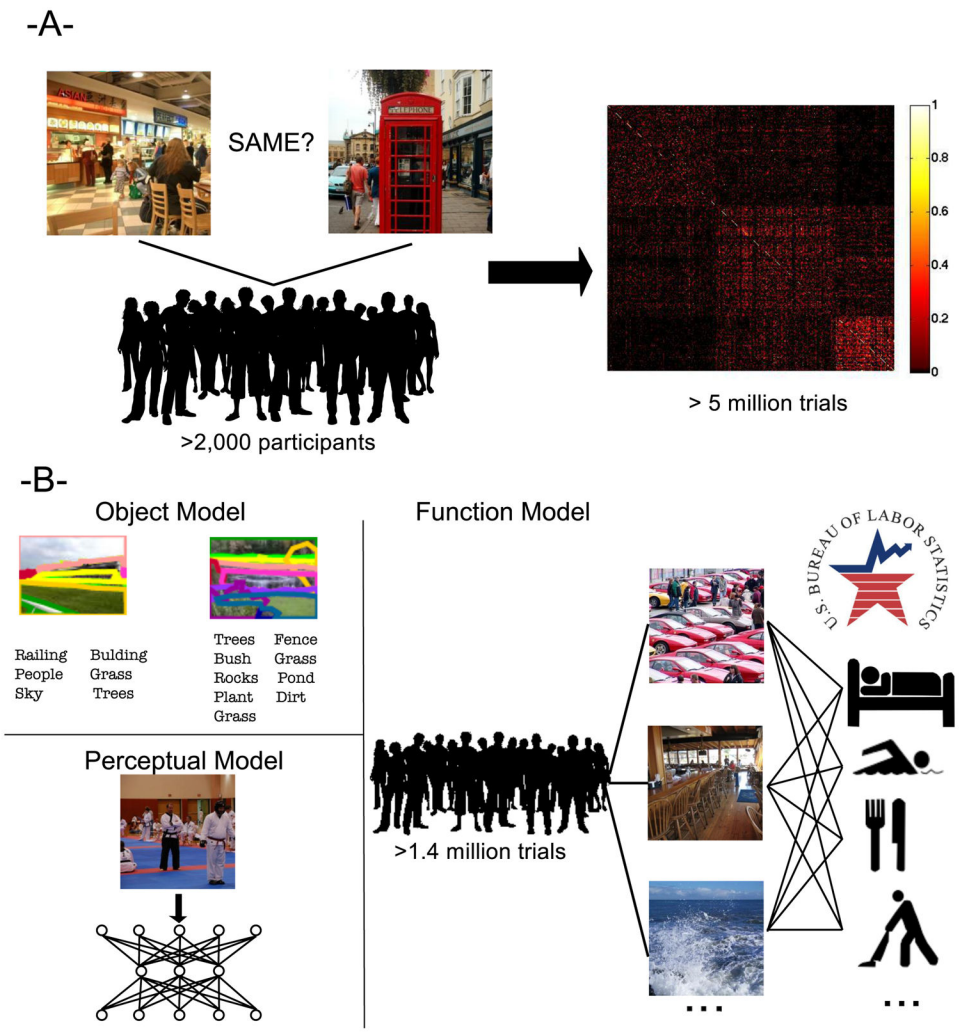

Figure 2.

(A) We used a large-scale online experiment to generate a distance matrix of scene categories. Over 2,000 individuals viewed more than 5 million trials in which participants viewed two images and indicated whether they would place the images into the same category. (B) Using the LabelMe tool (Russell, Torralba, Murphy, & Freeman, 2008) we examined the extent to which scene category similarity was related to scenes having similar objects. Our perceptual model used the output features of a state-of-the-art convolutional neural network (Sermanet et al., 2013), to examine the extent to which visual features contribute to scene category. To generate the functional model, we took 227 actions from the American Time Use Survey. Using crowdsourcing, participants indicated which actions could be performed in which scene categories.

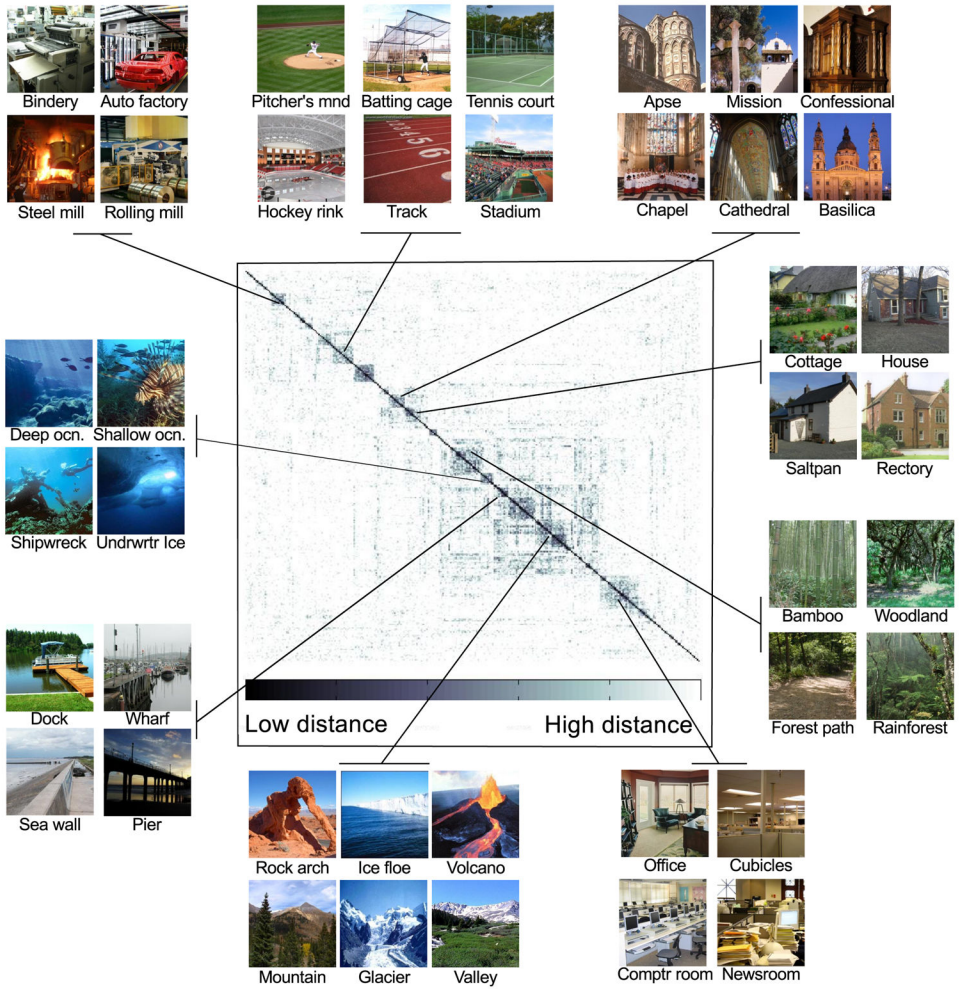

Figure 3.

The human category distance matrix from our large-scale online experiment was found to be sparse. Over 2,000 individual observers categorized images in 311 scene categories. We visualized the structure of this data using optimal leaf ordering for hierarchical clustering, and show representative images from categories in each cluster.

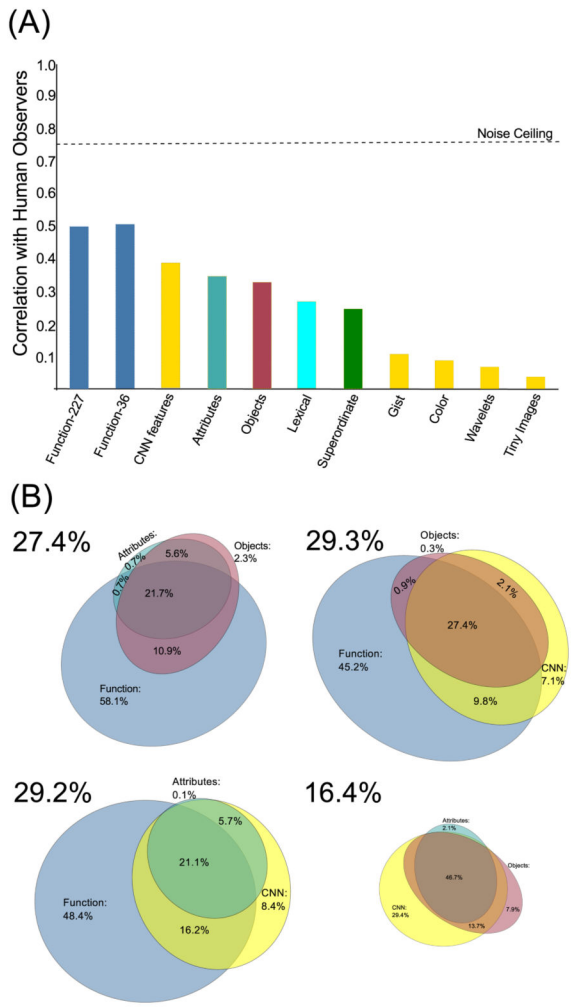

Figure 4.

(A) Correlation of all models with human scene categorization pattern. Function-based models (dark blue, left) showed the highest resemblance to human behavior, achieving 2/3 of the maximum explainable similarity (black dotted line). Of the models based on visual features (yellow), only the model using the top-level features of the convolutional neural network (CNN) showed substantial resemblance to human data. The object-based model, the attribute-based model, the lexical model and the superordinate-level model all showed moderate correlations. (B) Euler diagrams showing the distribution of explained variance for sets of the four top-performing models. The function-based model (comprehensive) accounted for between 83.3% and 91.4% of total explained variance of joint models, and between 45.2% and 58.1% of this variance was not shared with alternative models. Size of Euler diagrams is approximately proportional to the total variance explained.

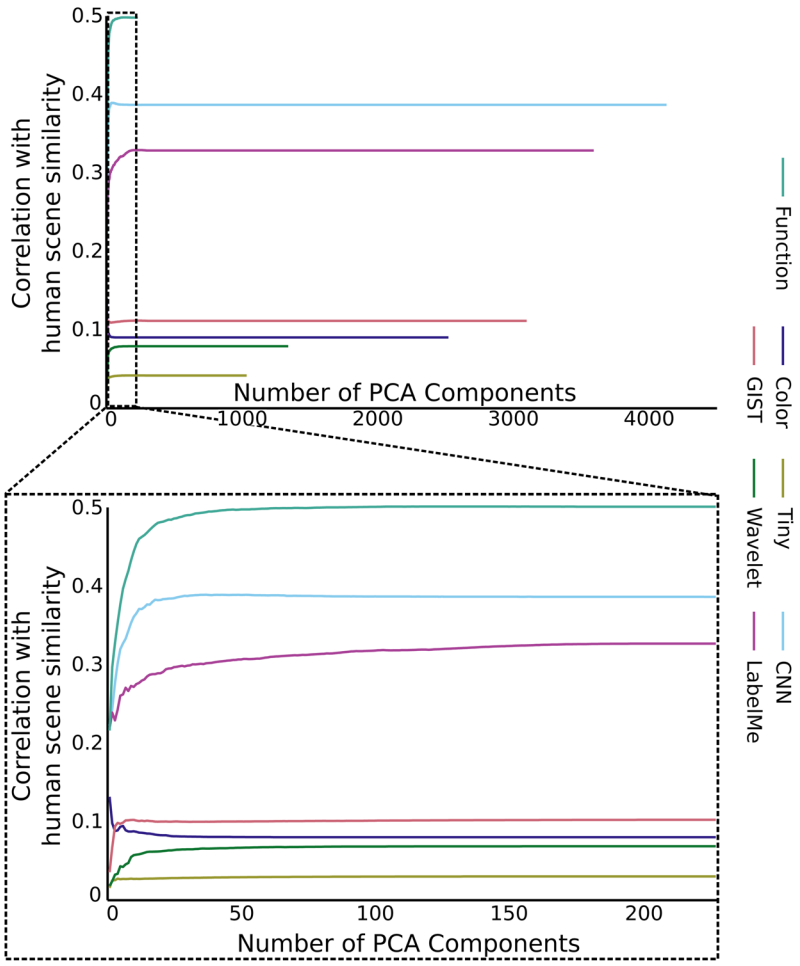

Figure 5.

Robustness to dimensionality reduction. For each feature space, we reconstructed the feature matrix using a variable number of PCA components and then correlated the cosine distance in this feature space with the human scene distances. Although the number of features varies widely between spaces, all can be described in ~100 dimensions, and the ordering of how well the features predict human responses is essentially the same regardless of the number of original dimensions.

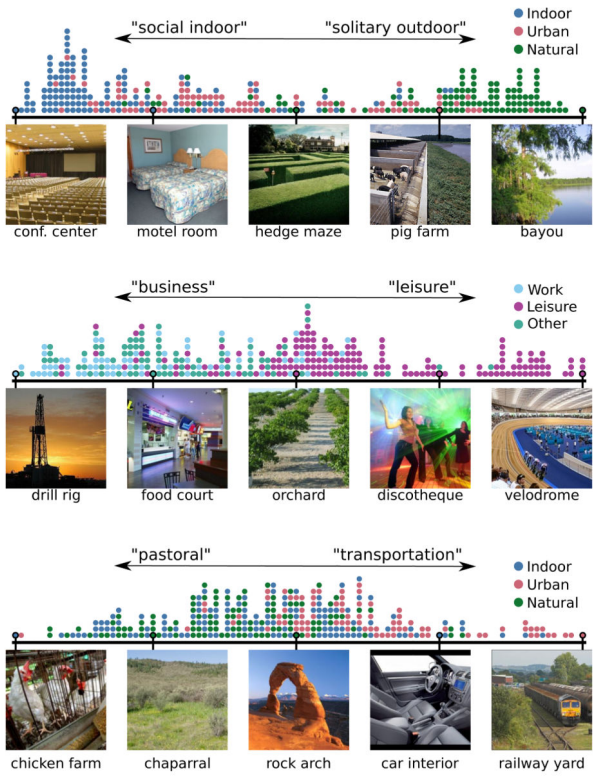

Figure 6.

(Top): Distribution of superordinate-level scene categories along the first MDS dimension of the function distance matrix, which separates indoor scenes from natural scenes. Actions that were positively correlated with this component tend to be outdoor-related activities such as *hiking* while negatively correlated actions tend to reflect social activities such as *eating and drinking*. (Middle) The second dimension seems to distinguish environments for work from environments for leisure. Actions such as *playing games* are positively correlated while actions such as *construction and extraction work* are negatively correlated (Bottom). The third dimension distinguishes environments related to farming and food production (pastoral) from industrial scenes specifically related to transportation. Actions such as *travel* and *vehicle repair* are highly correlated with this dimension, while actions such as *farming* and *food preparation* are most negatively correlated.

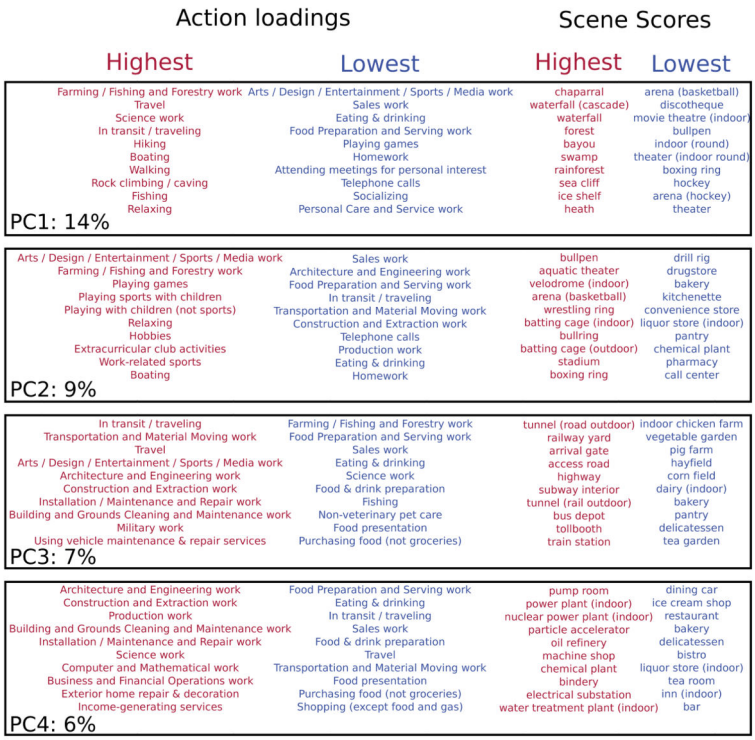

Figure 7.

Principal components of function matrix. MDS was performed on the scene by function matrix, yielding a coordinate for each scene along each MDS dimension, as well as a correlation between each function and each dimension. The fraction of variance in scene distances explained by each dimension was also computed, showing that these first four dimensions capture 81% of the function distance model.

Table 1

Variance explained (r^2) by fifteen regression models.

Model	r^2
Attribute	0.08
Object	0.11
CNN	0.15
Function	0.25
Object + Attribute	0.11
Attribute + CNN	0.15
Object + CNN	0.16
Object + Function	0.27
Attribute + Function	0.27
CNN + Function	0.29
Object + Attribute + CNN	0.16
Object + Attribute + Function	0.27
Attribute + CNN + Function	0.29
Object + CNN + Function	0.29
Attribute + Object + CNN + Function	0.29

Table 2

Correlation of top-four models in each of the three superordinate-level scene categories. The function-based model performs similarly in all types of scenes, while the CNN, attribute, and object-based models perform poorly in indoor environments.

	Indoor	**Urban**	**Natural**
Functions	0.50	0.47	0.51
CNN	0.37	0.43	0.59
Attributes	0.15	0.20	0.41
Objects	0.19	0.27	0.44