Estimating scene typicality from human ratings and image features

Krista A. Ehinger (kehinger@mit.edu)
Department of Brain & Cognitive Sciences, MIT, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

Jianxiong Xiao (jxiao@csail.mit.edu)
Computer Science & Artificial Intelligence Laboratory, MIT, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

Antonio Torralba (torralba@csail.mit.edu)
Computer Science & Artificial Intelligence Laboratory, MIT, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

Aude Oliva (oliva@mit.edu)
Department of Brain & Cognitive Sciences, MIT, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

Abstract

Scenes, like objects, are visual entities that can be categorized into functional and semantic groups. One of the core concepts of human categorization is the idea that category membership is graded: some exemplars are more typical than others. Here, we obtain human typicality rankings for more than 120,000 images from 706 scene categories through an online rating task on Amazon Mechanical Turk. We use these rankings to identify the most typical examples of each scene category. Using computational models of scene classification based on global image features, we find that images which are rated as more typical examples of their category are more likely to be classified correctly. This indicates that the most typical scene examples contain the diagnostic visual features that are relevant for their categorization. Objectless, holistic representations of scenes might serve as a good basis for understanding how semantic categories are defined in term of perceptual representations.

Keywords: scene perception; prototypes; categorization.

Introduction

Most theories of categorization and concepts agree that category membership is graded – some items are more typical examples of their category than others. For example, both sparrows and ostriches are birds, but a sparrow is generally regarded as a much more typical bird than an ostrich. The more typical examples of a category show many advantages in cognitive tasks. For example, typical examples are more readily named than atypical examples when people are asked to list examples of a category (eg., furniture) and response times are faster for typical examples when people are asked to verify category membership (eg., “a chair is a piece of furniture”) (Rosch, 1975).

According to Prototype Theory, concepts are represented by their most typical examples (Rosch, 1971). These prototypes are an average or central tendency of all category members. People do not need to actually encounter the prototypical example of a category in order to form a concept of that category; instead, they extract the prototype through experience with the variation that exists within the category (Posner & Keele, 1968).

Environmental scenes, like objects, are visual entities that can be organized in functional and semantic groups. Like other conceptual categories, scenes contain more and less typical exemplars. Tversky and Hemenway (1983) identified some typical examples of indoor and outdoor scene categories, but the total number of scene categories used in their studies was very small. Here, we extend the idea of scene typicality to a very large database containing over 700 scene categories. The goal of the current study is two-fold: first, to determine the prototypical exemplars that best represent each visual scene category; and second, to evaluate the performances of state-of-the-art global features algorithms at classifying different types of exemplars.

Method

Dataset

Stimuli were taken from the SUN Database, a collection of 130,519 images organized into 899 categories (see Xiao, Hays, Ehinger, Oliva & Torralba, 2010). This database was constructed by first identifying all of the words in a dictionary corresponding to types of places, scenes, or environments (see Biederman, 1987, for a similar procedure with objects). Our definition of a scene or place type was any concrete common noun which could reasonably complete the phrase, “I am in a *place*,” or “Let’s go to the *place*.” We included terms which referred to specific subtypes of scenes or sub-areas of scenes. However, we excluded specific places (like MIT or New York), terms which did not evoke a clear visual identity (like workplace or territory), spaces which were too small for a human body to navigate within (such as a desktop), and scenes with mature content. We included views of the interiors of vehicles (airplane cabin), but not exterior views of vehicles. We included specific types of buildings (skyscraper, house), because, although these can be seen as objects, they are known to activate scene-processing-related areas in the human brain (Epstein & Kanwisher, 1998). This procedure yielded an initial set of about 2400 scene words, and after combining synonyms and separating scenes with different visual identities (such as indoor and outdoor views), we obtained a list of about 899 unique semantic categories of scenes and places. For each of these categories, we collected a large set of images online, resulting in a database of about 130,000 images.

Note that there are different ways to define and categorize "scenes," which would generate a slightly different or more complete database than the one used here. For example, one might decide that different views of the same place qualify as different scenes, or one might choose to subdivide scenes based on spatial layout or surface features (e.g., forests with or without snow). However, this work represents the first attempt at estimating typicality on a dataset that is extensive enough to cover most of the plausible scene categories used to refer to places and scenes in discourse.

Stimuli

For the typicality experiment, we used the 706 scene categories from this database that contained at least 22 exemplars. Category size ranged from 22 images in the smallest categories to 2360 in the largest. A total of 124,901 images were used in the experiment.

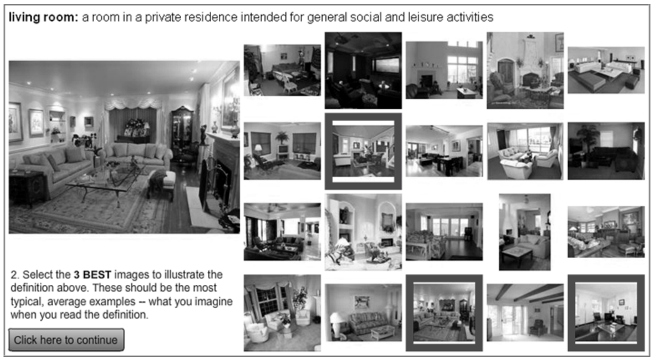
Participants

935 people participated in the experiment through Amazon's Mechanical Turk, an online service where workers are paid to complete short computational tasks (HITs) for small amounts of money. All workers were located in the United States and had a good performance record with the service (at least 100 HITs completed with an acceptance rate of 95% or better). Workers were paid $0.03 per trial.

Procedure

Participants were told that the goal of the experiment was to select illustrations for a dictionary. Each trial consisted of three parts.

First, participants were given the name of a scene category from the database, a short definition of the scene category, and four images. Workers were asked to select which of the four images matched the name and definition (one of the four images was drawn from the target category and the other three were randomly selected from other categories). The purpose of this task was to ensure that participants read the category name and definition before proceeding to the rating task.



living room: a room in a private residence intended for general social and leisure activities

2. Select the **3 BEST** images to illustrate the definition above. These should be the most typical, average examples – what you imagine when you read the definition.

Figure 1: The display seen by participants in the typicality rating task. In the experiment, images were shown in color.

Next, participants were shown 20 images in a 4 x 5 array. These images were drawn randomly from the target category, and did not include the image which had served as the target in the previous task. Images were shown at a size of 100 x 100 pixels, but holding the mouse over any image caused a larger 300 x 300 pixel version of that image to appear. An example of this display is shown in Figure 1. Workers were asked to select, by clicking with the mouse, three images that best illustrated the scene category.

In the third part of the task, workers were shown the same 20 images (but with their array positions shuffled) and were asked to select the three worst examples of the target scene category.

Design

On each trial, the set of 20 images was drawn randomly from the set of images in the target category. These random draws were such that each image appeared at least 12 times, and no more than 15 times over the course of the experiment. This resulted in 77,331 experimental trials. Each trial was completed by a single participant. Participants could complete as many trials as they wished; the mean number of trials completed per participant was 82.7 trials (median 7 trials).

Results

Participants' performance was judged on two measures: their performance on the 4AFC task, and whether they selected different images as the best and worst examples on a single trial. In general, participants performed well on the 4AFC task, with an average correct response rate of 97% (s.d. 0.13%). Most of the incorrect responses occurred on trials where one of the randomly-drawn foil images came from a category similar to the target category (for example, a cathedral might be the foil image for the category "basilica"). Participants were also reliably able to select different images as the best and worst examples of their category: participants marked an image as both best and worst on only 2% of trials (s. d. 0.10%); the likelihood of reselecting an image by chance is 40%. However, there were a few participants who reselected images at about chance rates, which suggests that they were selecting images at random with no regard for the task. We identified 19

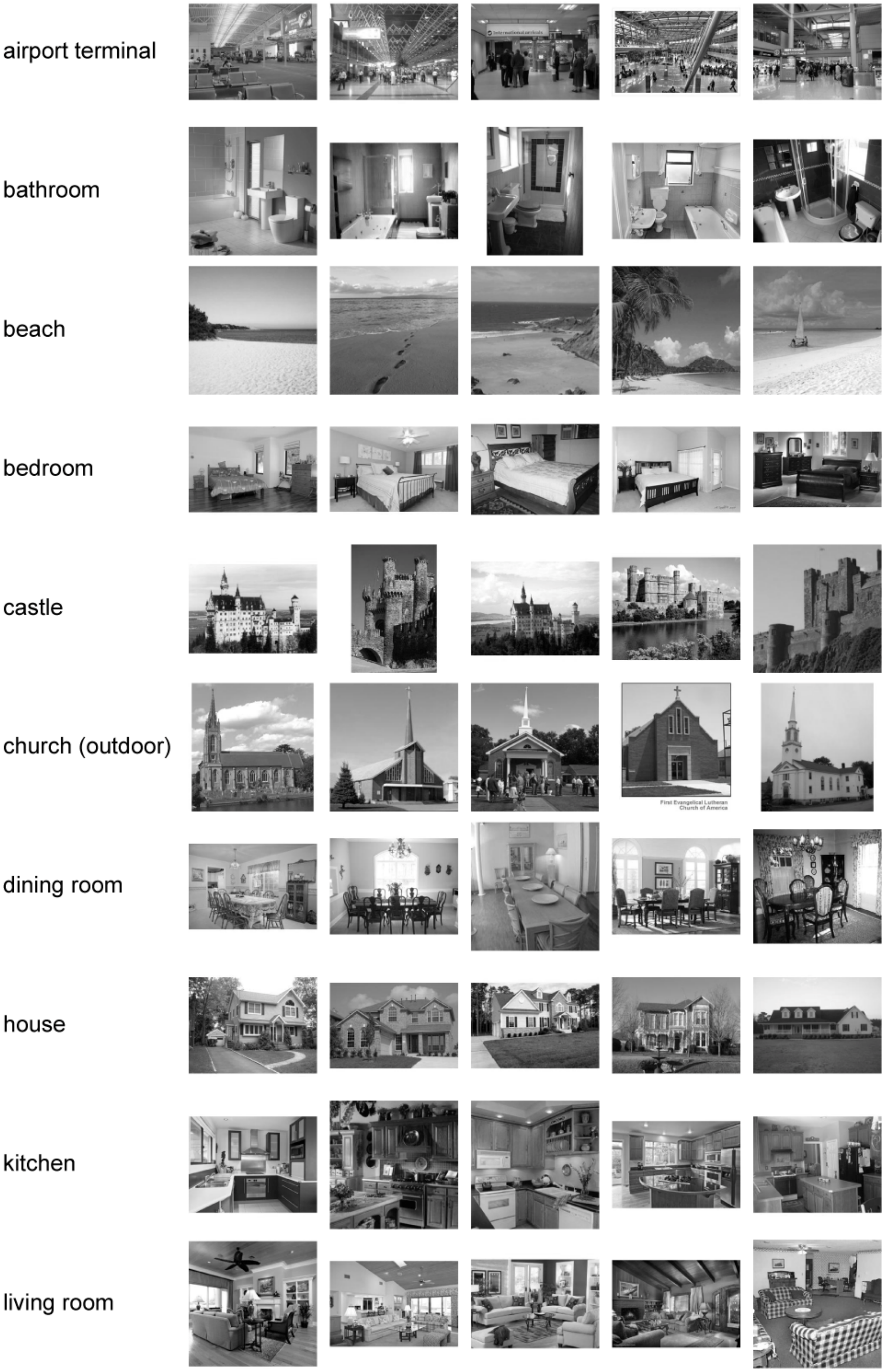


Figure 2: The five images rated most typical by participants, from the ten largest categories in the database.

participants who reselected the same images as both best and worst on at least 25% of trials (2% of total participants). Together these participants had submitted 872 trials (1.13% of trials), which were dropped from further analysis.

A "typicality score" was obtained for each image in the dataset. The typicality score was calculated as the number of times the image had been selected as the best example of its category, minus a fraction[1] of the number of times it was selected as the worst example, divided by the number of times the image appeared throughout the experiment. (Taking a fraction of the worst votes allows the number of "best" votes to be used as a tie-breaker for images that performed similarly.) A typicality score near 1 means an image is extremely typical (it was selected as the best example of its category nearly every time it appeared in the experiment), and a typicality score near -1 means an image is extremely atypical (it was nearly always selected as a worst example).

Examples of the most typical images from various categories are shown in Figure 2.

Comparison to chance

Even if participants selected "best" and "worst" examples at random, some images in each category would emerge as highly typical (or atypical) due to chance. It is important to check that the most typical images in this experiment are rated higher than would be expected if participants were simply responding randomly.

To check this, we ran a set of 100 simulations in which the images were rated randomly. Each image appeared in the simulation the same number of times it appeared in the actual experiment. On each appearance, the image had a 15% chance of being voted a "best" example, a 15% chance of being voted a "worst" example, and a 70% chance of receiving no vote. The simulation assumed that participants never selected the same image as both "best" and "worst" in a single trial, which was not actually true in the experiment. This means that the simulation actually overestimates the typicality scores that could be produced by random responses.

As shown in Figure 3, the typicality scores obtained by the most typical images in the experiment are much higher than the maximum scores that would be expected if participants were rating images randomly. More than half of the categories (401 out of 706) have a most typical image that is at least 3 standard deviations higher than the average "most typical image" from the simulation. This indicates that participants were selecting these images according to a strategy (such as selecting images that best matched their internal prototype for the scene category) and not just selecting images at random.

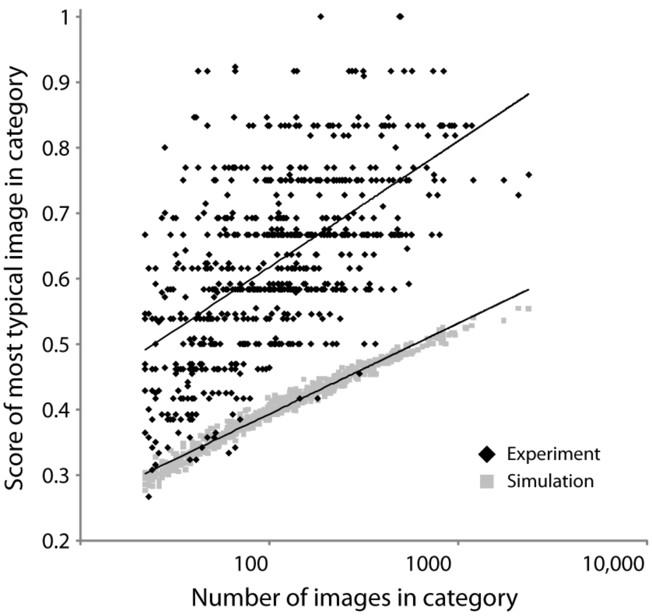


Figure 3: A comparison of the typicality scores obtained by the best image in each category in the expeirment to the typicality scores obtained in a simulation where images were rated randomly.

However, there are some categories where the most typical image scores no higher than would be expected from chance (46 categories have most typical image that is within a standard deviation of the average best score from the simulation). It's not clear why participants gave chance-like performance in these categories: although some of these categories are very unusual (e.g., rectory, cloister), many of them are familiar everyday categories like closet, desert road, or factory. This may reflect the distribution of exemplars we were able to obtain for these categories in the initial image search: it's possible that the images we collected for these categories were fairly homogeneous, with no particularly good or bad exemplars.

Typicality and models of scene classification

Do more typical exemplars of a scene category contain more of the visual features relevant to scene classification? To investigate this question, we classified scenes using the "all global features" classifier described in Xiao, et al. (2010). In computer vision, global features represent a class of algorithms that encode the spatial layout of textures in the image, without representing object information.

Global Features

As in Xiao et al. (2010), the all-feature kernel combines several representations that have been shown to be reliable for scene classification tasks. The GIST descriptor computes the output energy of 24 filters (8 orientations at 4 different scales) averaged on a 4x4 grid (Oliva & Torralba 2001). The Dense SIFT features (Lazebnik, et al., 2006) builds a coarse-to-fine spatial histogram pyramid of quantized

[1] This fraction was arbitrarily set to 0.9, but any value in the range 0.500 to 0.999 gives essentially the same results: changing this value changes the range of possible scores, but doesn't significantly change the rank order of scores within a category (90% of images move by less than 5 percentile points).

orientation histograms of image magnitude and orientation values on local patches. The HOG features (Dalal & Triggs, 2005; Felzenszwalb, et al., 2010) count occurrences of gradient orientation and use overlapping local contrast for normalization to improve invariance to changes in illumination or shadowing. While SIFT is known to be very good at finding repeated image content, the self-similarity descriptor (SSIM) (Shechtman & Irani, 2007) relates images using their internal layout of local self-similarities. Unlike GIST, SIFT, and HOG, which are all gradient-based approaches (measuring the density of the features), SSIM may provide a distinct, complementary measure of scene layout. Additionally, the "all-features" kernel includes histograms for specific geometric classes as determined by Hoiem et al. (2005), which represent aspects of a scene's spatial layout.

The "all global features" classifier is built from the large set of classifiers based on these state-of-the-art features. It covers a range of features which are likely to be important in scene recognition, including color histograms, representations of texture and scene regions (e.g., ground vs. sky), and information about edges and line orientations.

Classification procedure

Classifiers were trained with one-versus-all support vector machines as in Xiao et al (2010). In order to have enough exemplars for training and testing, the following simulations used the 397 categories that contain at least 100 exemplars. From each category, 50 images were selected at random to serve as the training set, and another 50 images were randomly selected to serve as the test set. Since the training and testing sets were chosen by random selection, they contained a range of more and less typical exemplars.

Xiao et al. (2010) found that the average performance of the "all global features" classifier on this 397-scene dataset is 38% (chance performance is 0.25%). What are the performances as a function of scene typicality? Figure 4 shows that classification of individual images varies with their typicality score: the most typical images were classified correctly about 50% of the time, and the least typical images were classified correctly only 23% of the time. Images were divided into four groups corresponding to the four quartiles of the distribution of typicality scores across the database. These groups contained 5020, 4287, 5655, and 4908 images (groups are listed in order from fourth quartile -- lowest typicality – to first quartile). A one-way ANOVA comparing these quartile groups shows a significant effect of image typicality quartile on classification accuracy (F(3,19846) = 278, p < .001); Bonferroni-corrected post-hoc tests show that the differences between each quartile are significant.

Image typicality is also related to the confidence of the SVM classifier. The confidence reflects how well the classifier believes the image matches its assigned category – scores near 1 indicate that the classifier is very confident that the image belongs in the category and scores near -1 indicate that the classifier does not believe the image belongs in the category. (Due to the difficulty of the one-versus-all classification task, confidence was low across all classifications, and even correctly-classified images had average confidence scores below zero.) Figure 5 shows the SVM confidence as a function of image typicality for correctly- and incorrectly-classified images. Confidence increases with increasing typicality, but this pattern is stronger in correctly-sorted images. A 4 x 2 ANOVA gives significant main effects of image typicality (F(3,19842) = 79.8, p < .001) and correct vs. incorrect classification (F(1,19842) = 6006, p < .001) and a significant interaction between these factors (F(3,19842) = 43.5, p < .001).

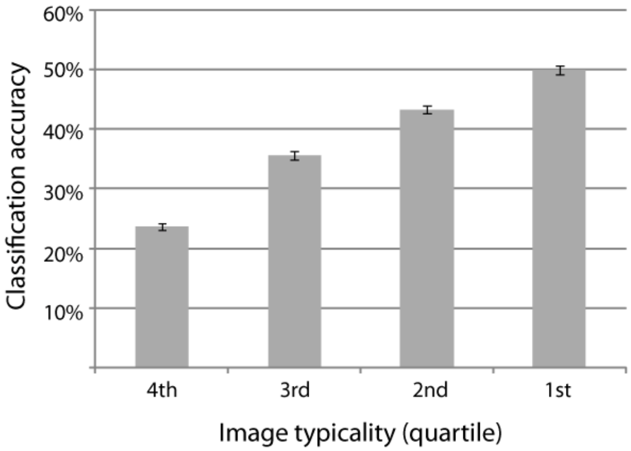


Figure 4: Performance of the SVM classifier as a function of image typicality. Images are sorted according to their typicality score from least typical (4th quartile) to most typical (1st quartile).

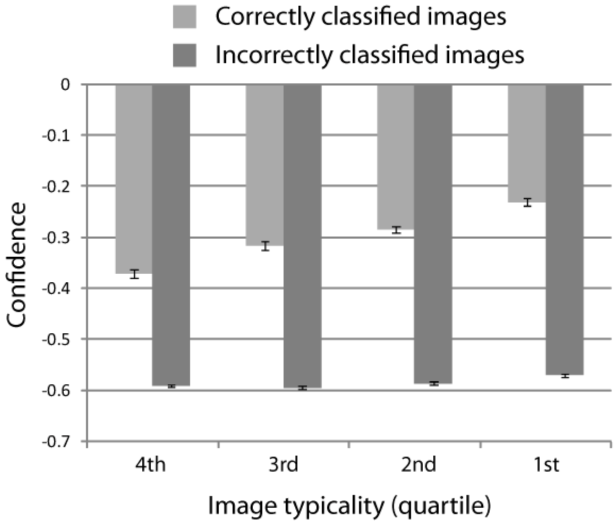


Figure 5: Confidence of the SVM classifier as a function of image typicality. Images are sorted according to their typicality score from least typical (4th quartile) to most typical (1st quartile).

Conclusion

Intelligent systems, artificial and biological, face the problem of how to organize complex stimulus representations. One framework for classifying scenes involves identifying visually informative features within a category. Previous attempts to characterize the categorical representation of environmental scenes have capitalized on uncovering a manageable set of dimensions, features, or objects with which to represent environments (Oliva & Torralba, 2001; Renninger & Malik, 2004; Fei-Fei & Perona, 2005; Lazebnik et al., 2006; Vogel & Schiele, 2007; Greene & Oliva, 2009; Ross & Oliva, 2010).

An alternate framework for classifying visual scenes appeals to their conceptual nature. Scenes, like individual objects, are associated with specific functions and behaviors, and have a categorical structure (Tversky & Hemenway, 1983). Here, we show that people have a representation of a typical or "best" exemplar for a wide range of scene categories. This elaborates on the scene prototype work of Tversky and Hemenway, and extends prototype research from the domains of objects, faces, and abstract patterns to scenes.

Furthermore, we show that scenes which people rate as more typical examples of their category are more likely to be correctly classified by computer vision algorithms based on global image features. Although we cannot claim that the features used in these algorithms are the same features which humans would use to perform the same classification task, this nevertheless indicates that more typical examples of a scene category contain more of the diagnostic visual features that are relevant for scene categorization.

Finally, this study is the first to show that reliable prototypes can be identified for a very large dataset of environmental scene categories, by both human observers and state of the art vision algorithms. One of the important distinctions between objects and scenes is that the categorical boundaries between scenes are less well defined than the boundaries between objects. Natural scenes in particular often lie on the boundary between two or more categories, like forest/mountain or river/lake (Vogel & Schiele, 2004), suggesting that typicality might be a particularly important concept for future progress in the field of human and computational scene understanding.

Acknowledgments

K.A.E. is funded by an NSF graduate research fellowship. This work is funded by NSF CAREER award (0546262), NSF grants (0705677 and 1016862), and a NEI grant (EY02484) to A.O; NSF CAREER award (0747120) to A.T; and Google research awards to A.O and A.T.

References

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review, 94,* 115-147.

Blanz, V., Tarr, M. J., & Bülthoff, H. H. (1999). What object attributes determine canonical views? *Perception*, *28*, 575-599.

Dalal, N., & Triggs, B. (2005). Histogram of oriented gradient object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 886-893.

Epstein, R. & Kanwisher, N. (1998) A cortical representation of the local visual environment. *Nature*, *392*, 598-60.

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *32*, 1627-1645.

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of Computer Vision and Pattern Recognition*, 524-531.

Greene, M.R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58*, 137-179.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of Computer Vision and Pattern Recognition*, 2169-2178.

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

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Renninger, L.W., & Malik, J. (2004). When is scene recognition just texture recognition? *Vision Research*, *44*, 2301-2311.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328-350.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

Ross, M.G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision, 10*, 1-25.

Shechtman, E., & Irani, M. (2007). Matching local self-similarities across images and videos. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition.*

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, *15*, 121-149.

Vogel, J., & Schiele, B. (2004). A semantic typicality measure for natural scene categorization. In *Pattern Recognition Symposium DAGM 2004*.

Vogel, J., & Schiele, B. (2007). Semantic model of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, *72*, 133-157.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN Database: Large scale scene recognition from abbey to zoo. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.