# Natural Scene Categorization from Conjunctions of Ecological Global Properties

**Michelle R. Greene (mrgreene@mit.edu)**
**Aude Oliva (oliva@mit.edu)**
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

**Abstract**

Human scene understanding is remarkable: with only a brief glance at an image, an abundance of information is available - spatial layout, scene function, semantic label, etc. Here we propose a scene-centered model of rapid human scene understanding that uses a vocabulary of global, ecological scene properties that combine to categorize natural landscape images. Behaviorally, we show human observers are sensitive to the underlying distributions of these global properties for use in basic-level categorization. An ideal observer trained only on the distributions of these properties predicts human scene categorization performance (r=0.90) and human errors.

## Introduction

Human scene understanding is truly remarkable: with the briefest of glimpses at an image, we instantaneously understand its content and meaning (Potter, 1975; Thorpe et al., 1996). Even more striking is the richness of the variety of information perceived within a glance: a few objects, spatial layout, functional and conceptual properties and even emotional valence (Maljkovic and Martini, 2005) are all available with well under 100 msec of exposure to a novel image. The entirety of this information is termed a scene's *gist* (Oliva, 2005). What is the nature of the representation that mediates rapid scene categorization?

To the contrary of the traditional ideas of research in scene understanding that treat objects as the atoms of recognition, we consider that real world scenes can be recognized without necessarily identifying the objects they contain (Biederman et al, 1982; Greene and Oliva, 2005; Schyns & Oliva, 1994; Oliva & Schyns, 2000). This *scene-centered approach* to recognition emphasizes properties describing the structure and the meaning of the whole scene independent of object analysis. Recent computational models of scene recognition have shown indeed that a variety of low level features (color, texture) and spatial layout properties (e.g. its level of openness, perspective) are correlated with the semantic category of environmental scenes at both superordinate and basic level of representation (Fei Fei & Perona, 2005; Oliva & Torralba, 2001; Walker-Renninger and Malik, 2001; Torralba & Oliva, 2003; Vogel & Schiele, 2004). A scene-centered schema would not preclude local object recognition, but would serve as a feed-forward and parallel pathway of visual processing, enabling the rapid estimation of scene gist.

The behavioral and modeling experiments we propose here are meant to establish the psychological foundation of a scene-centered approach to scene understanding. Beyond the principle of recognizing the "forest before the trees" (Navon, 1977), we propose an operational definition of the global scene properties permitting the categorization of a scene as a "forest". Faithful to a scene-centered representation which will capture the completeness of the gist of a scene, our selection of a vocabulary of global scene properties was influenced by the requirement to describe structural, functional and surface-based features of an environmental scene. Namely, which properties of a space allow the description of its semantic category, function and affordance?

Previous research has shown that global properties of *mean depth*, *openness* and *expansion* describe the spatial layout of a scene well enough to be predictive of its probable semantic category (Oliva & Torralba, 2001). Properties of *navigability* and *camouflage* reflect the functionality of the space and the type of actions that can be afforded in outdoor natural scenes. *Movement* (i.e. the transience of the elements in the scene) and *temperature* are relevant surface-based properties that influence human's behavior, and refer to the material and texture qualities of image regions (i.e. rocky and sandy often imply *hot* and *non-moving*, while snow implies *cold* and rushing water implies *movement*). These properties have been shown in previous work to be available for report with less exposure time than the semantic category of an image (Greene & Oliva, 2005).

The seven global properties we describe here are *ecological* in the sense that they are descriptive of the types of interactions a human could have in an outdoor natural landscape (e.g. can walk through without worry of occluding objects), or are descriptive of the space of a scene (e.g a panoramic environment), which can in turn, guide behavior. It is of note that such a scene-centered representation has no explicit declaration of objects or region segmentation. Outdoor scenes have few objects that can be manipulated and interacted with by a human (e.g. a rock, a flower), but their size is almost entirely local and therefore not captured by global properties.

Our principal hypothesis is that the initial image representation that facilitates semantic scene categorization can be built from the conjunctive detection of ecological

global properties. In the following, we evaluate the extent to which global properties uniquely describe the basic-level category of natural scenes (Experiment 1). Then, we show the causal relationship existing between global properties and rapid categorization (Experiment 2). Finally, we demonstrate that an ideal observer model whose only access to scene information is through global properties can predict human rapid categorization performance of natural scenes. All together, these results provide support for an initial image representation that is scene-centered, global and explicitly representing scene function.


Figure 1: Examples of scenes images ordered along four global properties. For each property, a scene image with a low, medium and high magnitude is shown.

## Experiment 1: Norming study

The goal of the first experiment was to obtain a measure of the magnitude of each global property in 200 images depicting a variety of natural landscapes. First, the images in the database were selected as prototypical examples of one of the following eight categories: desert, field, forest, lake, mountain, ocean, river and waterfall (with 25 images per category) by three independent observers. Next, we obtained rankings on each scene's degree of openness, camouflage, navigation, etc. Figure 1 illustrates low, medium and high magnitude examples of four global properties. Fifty-five observers (25 males, mean age 28) with normal or corrected-to-normal vision, consented to rank the 200 pictures for monetary compensation. These rankings served as ground truth for image selection in Experiment 2 as well as training information for the model observer.

The magnitude measures were obtained using a hierarchical grouping procedure (Oliva & Torralba, 2001). First, 100 picture thumbnails appeared on an Apple 30" monitor (size of 1.5 x 1.5 deg / thumbnail), placed in a 10 x 10 grid. The interface allowed participants to drag images with a mouse to one side of the screen or the other and view a larger version of the image by double-clinking on the thumbnail. Participants were instructed to divide the images into two groups based on a specific global property, such that, for example, images with a high degree of this property (e.g. openness) were on the right side of the screen and images with a low degree of openness on the left side. In a second step, participants were asked to split each group into two finer divisions. Finally, the groups were split again to form a total of 8 groups, ordered from the highest to the lowest magnitude for a given property. At any point during the trial, participants were allowed to move an image to a different subgroup, to refine the ranking, and participants had unlimited time to perform this task. Participants repeated this hierarchical sorting process on the remaining 100 pictures in database along the specified global property. Each participant ranked the image database on one or more global properties such that each global property was finally ranked by ten participants. The global properties were described as follows:

**Camouflage:** How efficiently and completely could you hide in the environment? The possibility for camouflage ranges from complete exposure in an environment (no place to hide) to completely concealable due to dense foliage, etc.
**Movement:** At what rate is the scene moving or changing? This can be related to actual physical movement such as a running river, or the transience of the scene (the sun setting, the fog lifting, etc.) At one extreme, a scene will only be changing in geological time and at the other extreme, the gist of the picture depends on it having been taken at that moment.
**Navigation:** How easy or difficult would it be for a human to traverse the environment from the given viewpoint to the

horizon? This ranges from complete impenetrability of the environment to a human trying to walk through to the possibility of walking nearly infinitely in any direction without obstacle. Navigable images are not necessarily low camouflage as there can be a clear path through a dense forest, for example.

**Temperature:** What is the physical temperature of the outdoor environment? This ranges from the coldest place to the hottest place.

**Openness:** Is there a clear view to the horizon line? At one extreme, there is no horizon line or visible sky and the scene is entirely enclosed, and at the other, there is a clear definable horizon in the middle of the image. Openness is a property of the viewpoint of the image, and is therefore not correlated with camouflage.

**Expansion:** Is there perspective in this image with converging parallel lines, or is the viewpoint flat on a single surface? Although somewhat correlated with navigability (e.g., many roads show strong linear perspective), expansion describes the space of the environment independently of its affordances.

**Depth:** What volume does the scene subtend? Is it a close-up shot from 1 meter away, or is it a panorama where one can see for miles? A scene may have large volume independent of other spatial layout, interactive or surface properties.

**Results**

There was strong agreement among participants for global property rankings: between-observer Spearman's rank-correlations ranged from 0.6 (movement) to 0.83 (openness), and were all statistically significant ($p < .01$). This indicates that participants ranked the same images in very similar ways, suggesting that these properties correspond to objective interpretations of the image.

The mean magnitude rank for each semantic category along the seven global properties is shown in Figure 2. Interestingly, we observed that the distribution of global property magnitudes provide a unique description of each basic-level category. Some categories such as *lake* or *mountain* have equal weights for all global properties, whereas other categories such as *desert*, *waterfall*, and *forest* have properties that are clearly diagnostic (shown by high and low peaks). The set of magnitudes represents the average exemplar of a given category: for instance, a *desert* is a very *hot* and *open* environment, with low degree of movement and camouflage; *waterfall* and *river* have a high degree of movement (due to rushing water); *forests* are *closed* environment with a high potentiality for *camouflage*.

The results suggest that the global properties constitute a conceptual signature of the meaning of a specific basic-level natural category and suggest the possibility that scene understanding may be built upon these global signatures, a hypothesis we further investigate in Experiment 2.

## Experiment 2

According to a scene-centered approach to image understanding, the semantic category can be represented as a conjunction of global properties, describing diagnostic information about the scene spatial layout and its functional properties. Here we test the extent to which global property information is used by people in rapid scene categorization.

From the ranking study, we know that particular magnitudes of global properties are diagnostic for certain semantic categories (e.g. high temperature is a robust regularity in deserts). We reason that if global property information is being used by human observers to identify the scene category, then presenting images from one category among distractors from other categories but with a similar global property magnitude (e.g. a hot beach scene) should lead to more false alarms in a yes-no forced choice categorization task.

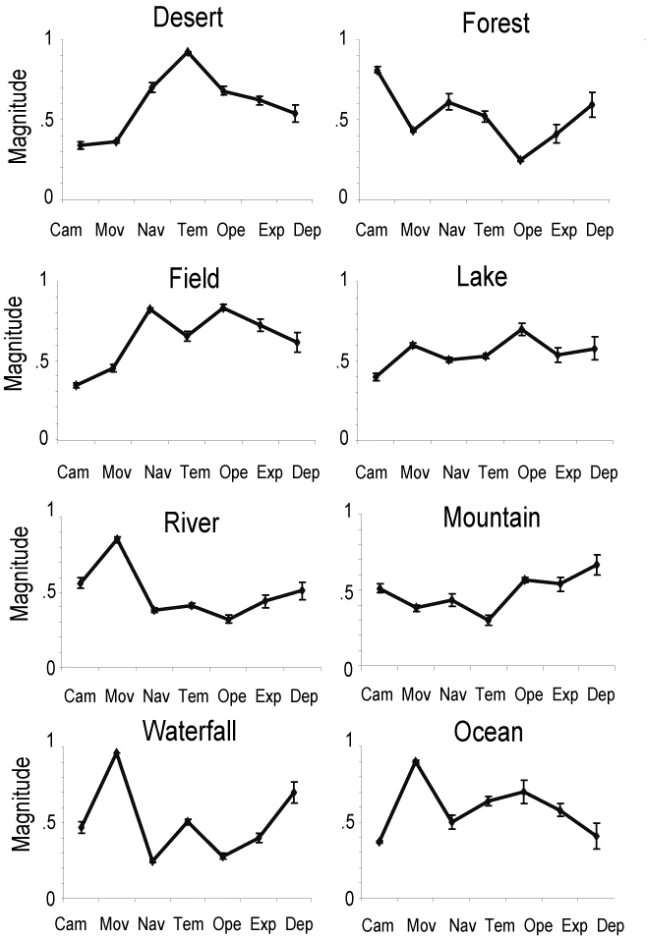


Figure 2: Mean magnitude of each global property, for each scene semantic category. (Cam= camouflage; Mov = movement, Nav=navigation, Tem = temperature, Ope = openness, Exp = expansion, Dep = mean depth).

## Method

Thirty-two observers, with normal or corrected to normal vision, participated in Experiment 2 (11 males, mean age 22.4). Participants were given the name of a category and asked to answer as quickly and as accurately as possible whether the briefly presented full-color picture (30 msec duration followed by a 1/f noise mask) belonged to the target category. The procedure consisted of a full confusion matrix of experimental blocks, where each target category was compared to distractor sets with particularly "high" or "low" magnitudes on one of the seven global properties, yielding 112 conditions (8 target categories * 7 global properties * 2 magnitudes). For instance, if "forest" was the selected target category, pictures of forests would be categorized among distractors from images from a variety of semantic categories but who shared a particular global property magnitude, such as "high movement". Each individual completed at least 8 blocks that were diagonalized such that no participant saw the same distractor set twice. Each experimental block was composed of 25 target images and 25 distractor images and participants were told to answer as quickly and as accurately as possible whether the briefly presented scenes belonged to the target category by pressing a 'yes' or 'no' key. Finally, each of the 112 experimental blocks was completed by six meta-subjects.

### Results

As expected, human hit performances on all categories was high: ranging from 0.72 for oceans to 0.90 for forest and 0.94 for waterfalls.

We analyzed the false alarms for the confusion matrix, comparing them to errors predicted from the ranking experiment. For each category, predicted false alarms for the confusion matrix were created by expressing the mean global property magnitude values as a distance from the mean values of these properties for all categories. From Figure 2, this corresponds to the absolute magnitude difference of each property from the 0.5 level. The greater this number, the more diagnostic a property is for this category (for instance, high camouflage for forest). Again, we predict that the normalized false alarm rate will be highly correlated with this measured diagnosticity value.

Figure 3 shows that the normalized false alarm rate for the confusion matrix is significantly correlated with global property diagnosticity ($r=0.47$, $p=0.0003$). Red bars going below the 0.5 line indicate an increase in false alarms in the direction of the low magnitude end of the global property, whereas bars above this line indicate false alarm increases towards the high end. Different distractor sets produced radically different false alarm rates, even within the same category. This result indicates that human observers are sensitive to a category's distribution of global properties, and use this information to aid rapid categorization.

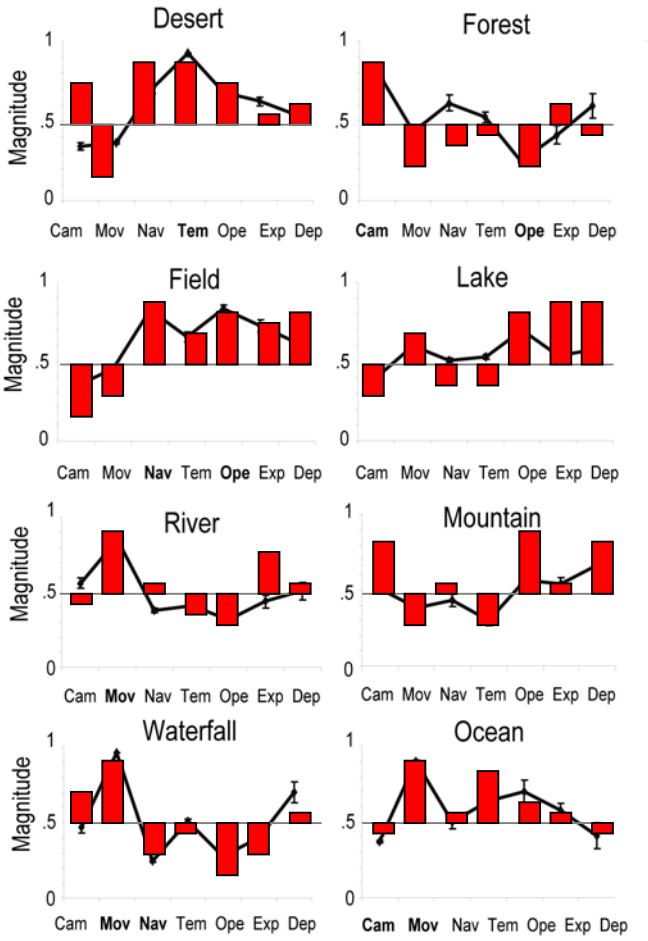


Figure 3: False alarms (in % above category baseline, shown in red), are significantly correlated with predictions made from the ranking experiment, indicating that global property information for a category is weighted in a rapid categorization task proportional to how it is diagnostic of the category.

## Ideal observer model

Experiments 1 and 2 have shown that global property information is useful to humans in rapid categorization tasks. We next asked: to what extent can human performances be predicted using *only* global property information? To test this, we built an ideal observer model to do this task. While most ideal observer analyses examine how close human observers are to the mathematical optimum for a given task, ideal observers have also been used to test hypotheses about perceptual mechanism (Geisler, 2003). Here we test the hypothesis that scene categorization can be done by conjunctions of global properties by building a conceptual ideal observer whose only information about scene categories is from the categories' distributions of global properties.

Using the global property rankings to train the model, we ran the model 25 times, testing each image in turn. In each run, 24 images from each semantic category (192 total) served as training, and the last eight (one from each category) were used for testing. The observer was given the semantic category labels for each of the training images, and computed the mean and variance along each of the global properties for each category.

In testing, the model was presented with the global property descriptors of the eight test images. The model computed the maximum likelihood category ($h_{ML}$) for this image given the distributions of global properties learned in training.

$$h_{ML} = \arg\max_{h \in H} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

## Results

The ideal observer's categorization performance (hits) was remarkably similar to that of the human observers in the behavioral experiment (r=0.90, p=0.0002, Figure 4).

### Error analysis

Does the ideal observer make the same kinds of errors that human observers do? There was a significant correlation between the number of false alarms made to an image by human observers and failure of the ideal observer (r=0.66, p=0.001). Furthermore, the nature of the errors was highly similar. Given an error of the ideal observer (i.e. outputting that an image is a *lake* when it is really an *ocean*), human observers made the same mistake in 69% of the images. (Chance is 12.5%). Examples of the correct responses and the false alarms made by the model and/or human observers are shown in Figure 5. Figure 5a shows images well-classified by both human and the model. Some images are not well classified by either (Figure 5b), and seem to correspond to less prototypical instances of the scene category Figure 5c and 5d show examples of images classified incorrectly by humans, but not by the ideal observer model, and vice versa.

### Diagnosticity of global properties for model

The ideal observer shows that these seven global properties are sufficient to predict human performance in a rapid scene categorization task. However, it does not indicate whether all of the properties are necessary for categorization. To test, we compared the confusion matrix of human categorizations to runs of the ideal observer model that was trained without a particular global property. Both of these "knock-out" a global property for use in categorization: for the humans, as the distractor set had a uniform distribution for this global property, it cannot inform categorization. In other words, assuming *movement* is diagnostic of *ocean*, classifying oceans among high movement distractors will render movement useless for the task. For the model, the global property is knocked out because there is no representation of the property at all.

For the ideal observer, knocking out any global property significantly decreased the model's categorization performance to a similar degree (hit rate decreasing from mean of 74% to a mean of 67%). However, each global property had unequal contributions across categories. Each category had a unique set of necessary global properties.

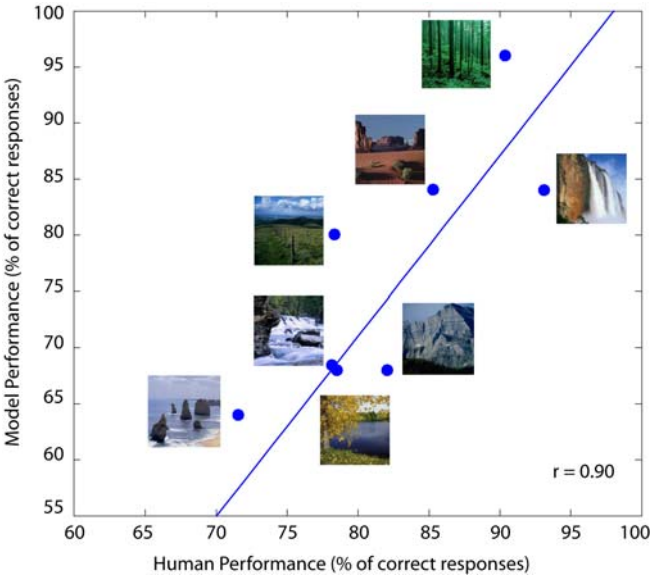


Figure 4: Ideal observer categorization performance (hits) is well-correlated with human rapid categorization performance. Scene categories that are well-classified by humans are well-classified using only global property information.

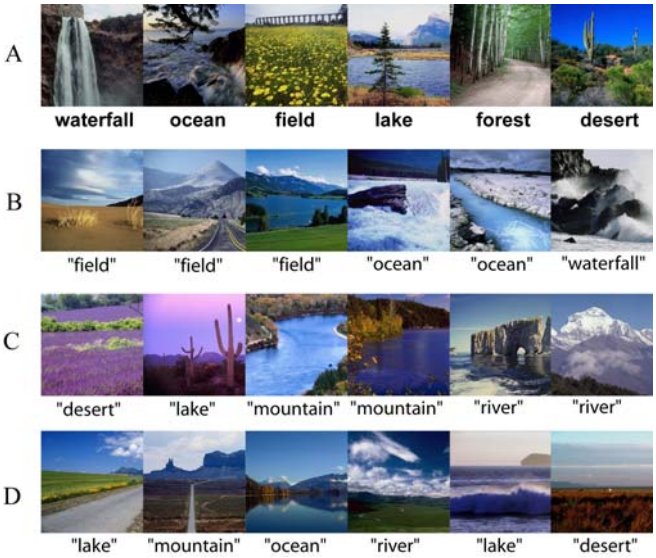


Figure 5: A (bold titles) corresponds to the correct responses made by both humans and the ideal observer model for the above scene pictures. The other rows (with titles in quotes) represent categorization errors made respectively by both humans and the model (B); by the model only (C); by the humans only (D), for the respective scene pictures.

In addition, for both the human and ideal observers, we converted false alarm rates into percent increases in false alarms over the baseline false alarm rate for the given category (as some categories are intrinsically more difficult than others). The correlation between human and model false alarms was 0.83 ($p<0.0001$), indicating that human and ideal observers are impaired by the loss of particular global properties for categorization and suggesting that the information used by both observers might be the same.

### Discussion

In this work, we have shown that a scene-centered approach to image understanding predicts human rapid scene categorization. Our approach uses a short vocabulary of global and ecological scene properties that combine to categorize a variety of natural landscape environments. In this work, we have shown that human observers classify images as points along global property dimensions in a consistent way (Experiment 1), and that information from these properties is weighted in rapid categorization tasks in a way that follows the distribution of the properties' regularities in the database (Experiment 2). Finally, we have shown that a model can predict human performance in terms of accuracy and error type with *only* information from these global properties.

It has been known for some time that visual perception tends to proceed in a global-to-local manner, but for stimuli as complex as a natural image, it is not immediately obvious what the nature of the global features are. By grounding our search in the principles of environmental affordance (Gibson, 1979; Rosch, 1978), we have been able to find a collection of properties that are necessary and sufficient to capture the essence of many landscape image categories. These global properties are also unique in the sense that they span other types of scene descriptors such as spatial layout (openness, expansion and mean depth), function (camouflage and navigability) and surface type (movement and camouflage). However, all of these are ecological because layout and surfaces also guide the types of action (or affordances) of the environment.

All together, our results provide support for an initial scene-centered visual representation used by human observers, and built on conjunctions of global properties that explicitly represent scene function and spatial layout.

## Acknowledgments

The authors wish to thank George Alvarez, Barbara Hidalgo-Sotelo, Todd Horowitz, Talia Konkle, John Kraemer, Mary Potter, Josh Tenenbaum, Antonio Torralba Jeremy Wolfe and three anonymous reviewers for helpful comments on the research and the manuscript. This research is supported by a NSF Graduate Research Fellowship awarded to MRG, and a NEC Corporation Fund for Research in Computers and Communications awarded to A.O.

## References

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, l4, l43-l77.

Geisler, W.S. (2003) Ideal Observer Analysis. In: L Chalupa and J. Werner (Eds.) *The Visual Neurosciences.* Boston: MIT press, 825-837

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton-Mifflin.

Greene, M. R., and Oliva, A. (2005) Better to run than hide: The time course of naturalistic scene decisions, *J. Vis.*, 5, 70a.

Maljkovic, V., and Martini, P. (2005) Short-term memory for scenes with affective content. *J. Vis.*, 5: 215-229.

Navon, D. (1977) Forest before trees: the precedence of global features in visual perception. *Cognit. Psychol*.,9: 353-383.

Oliva, A. (2005) Gist of the Scene. In Neurobiology of Attention, L. Itti, G. Rees and J. K. Tsotsos (Eds.), Elsevier, San Diego, CA (pp 251-256).

Oliva, A., and Schyns, P.G. (2000) Diagnostic colors mediate scene recognition. *Cognit. Psychol*., 41: 176-210.

Oliva, A., and Torralba, A. (2001) Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *Int. J. Comp. Vis*.,42: 145-175.

Potter, M.C. (1975) Meaning in visual scenes. *Science*, 187, 965-966.

Rosch, E. (1978). Principles of categorization. In: E. Rosch, B. Lloyd (eds.): Cognition and categorization. Hilldale, NJ: Lawrence Erlbaum.

Schyns, P.G., and Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychol. Sci.*, 5, 195-200.

Thorpe, S., Fize, D., and Marlot, C. (1996) Speed of processing in the human visual system. *Nature*, 381: 520-522.

Torralba, A., and Oliva, A. (2003) Statistics of Natural Images Categories. *Network: Computation in Neural Systems*, 14, 391-412.

Vogel, J., and Schiele, B. (2004) A Semantic typicality measure for natural scene categorization. *Proc. Of Pattern Recognition Symposium DAGM*, Tubingen, Germany.

Walker Renninger, L., and Malik, J. (2004) When is scene identification just texture recognition? *Vision Res*., 44: 2301-2311.