A Global Framework for Scene Gist

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY
OCT 0 6 2009
LIBRARIES

By

Michelle R. Greene

B.S., Psychobiology
University of Southern California, 2004

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COGNITIVE SCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SEPTEMBER 2009

© 2009 Massachusetts Institute of Technology. All rights reserved.

ARCHIVES

Signature of author__
Department of Brain and Cognitive Sciences
08-13-2009

Certified by____________________
Aude Oliva
Associate Professor of Brain and Cognitive Sciences
Thesis supervisor

Accepted by____________________
Earl K. Miller
Picower Professor of Neuroscience
Chairman, Committee for Graduate Students

A Global Framework for Scene Gist
by
Michelle R. Greene
Submitted to the Department of Brain and Cognitive Sciences in
Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Cognitive Science

Abstract

Human observers are able to rapidly and accurately categorize natural scenes, but the representation mediating this feat is still unknown. Here we propose a framework of rapid scene categorization that does not segment a scene into objects and instead uses a vocabulary of global, ecological properties that describe spatial and functional aspects of scene space (such as navigability or mean depth). In Chapter 1, four experiments explore the human sensitivity to global properties for rapid scene categorization, as well as the computational sufficiency of these properties for predicting scene categories. Chapter 2 explores the time course of scene understanding, finding that global properties can be perceived with less image exposure than the computation of a scene's basic-level category. Finally, in Chapter 3, I explore aftereffects to adaptation to global properties, showing that repeated exposure to many global properties produces robust high-level aftereffects, thus providing evidence for the neural coding of these properties. Altogether, these results provide support for the hypothesis that rapid categorization of natural scenes may not be mediated primarily though objects and parts, but also through global properties of structure and affordance.

Thesis supervisor: Aude Oliva
Title: Associate Professor of Brain and Cognitive Sciences

Table of Contents

Acknowledgements

First, I would like to thank my advisor, Aude Oliva. Throughout my graduate school experience, she has given me a great deal of intellectual independence as well as focused attention. She has also served as a strong model of creativity and bravery in science. I sincerely appreciate her focus on big and new ideas, and look forward to being her colleague in the future.

I would also like to thank my thesis committee, Molly Potter, Jeremy Wolfe and Ted Adelson for their continued support of my research, their helpful comments and suggestions, and for all providing me examples of what it means to do interesting and rigorous research. Although not an advisor in a formal sense, I would also like to thank Antonio Torralba for his continued advice on these projects.

I would also like to thank the members of the BCS faculty who have shaped and influenced by thinking outside of a direct advisory role: Nancy Kanwisher, Josh Tenenbaum, Pawan Sinha and Ruth Rosenholtz. Most sincere thanks go to Denise Heintze and Bettiann McKay whose talents allow science to happen.

CVCL is a lab of diverse backgrounds and interests, and I have benefited enormously from the help of all lab members, former and current: George Alvarez, Tim Brady, Emmanuelle Boloix, Krista Ehinger, Olivier Joubert, Talia Konkle, Soojin Park, and Michael Ross. I must single out my "litter mate", Barbara Hidalgo-Sotelo out for special thanks for her kindness, grace, tenacity and so many other fine traits that are too numerous to list here.

I would also like to thank other friends who have influenced my life in scientific and recreational matters: Liz Bonawitz, Adam Davis, Mike Frank, John Kraemer, Elizabeth Levy, Yuanzhen Li, Vikash Mansinghka, Retsina Meyer, Ethan Meyers, Dominique and Tracy Pritchett, Lauren Schmidt, Lavanya Sharan, Todd Thompson, Ed Vul and Brad Wyble.

Finally, I would like to thank my family because without them, I cannot imagine being here today. My parents, Kenneth and Carolyn Greene, have always provided me with unconditional, loving support throughout my education. They are true role models of courage, determination, grit, endurance and compassion. I am honored to call them friends as well as family. My "kid" brother Eric also has taught me that one can be both accomplished and balanced. Last, thanks go to my partner, Adam Katz, who has been amazingly supportive and loving. He has provided encouragement, skepticism, a listening ear and hugs all in the right proportions and at the right times.

Chapter 1

Introduction

When we look around our environment, our gaze shifts rapidly, providing us with about three images a second (Rayner, 1998). While many of these visual scenes may be familiar, we have no trouble recognizing the ones that are completely novel, such as entering a friend's house for the first time, or walking down the street in a new city. Indeed, numerous laboratory studies have shown that human observers can glean a great deal of information from a single fixation on a novel scene: understand its semantic topic or category (e.g. "birthday party": Intraub, 1981; Potter, 1975); determine whether the scene is natural or urban (Joubert, Rousselet, Fize & Fabre-Thorpe, 2007; determine the presence of a large object (Thorpe, Fize & Marlot, 1996; Van Rullen & Thorpe, 2001); or even describe how pleasant the scene is (Kaplan, 1992). However, we do not yet know how such a rich representation is built so quickly by the visual system.

Two types of initial scene representations have been proposed by the literature. The first asserts that a scene can be understood from the identification of the objects it contains. One might do this by recognizing a particularly prominent or diagnostic object such as a refrigerator in a kitchen scene (Friedman, 1979), or by recognizing a few objects that are contextually related to the scene category and arranged in a spatially stereotyped manner, such as a desk, chair and computer monitor in an office (Biederman, Blickle, Teitelbaum, Klatsky, & Mezzanotte 1988). The second type of initial scene representation is a global pathway in which features from the whole scene allow the recognition of the place and subsequent recognition of the objects within the scene. Global scene features might include the aggregate shape of an arrangement of smaller elements or objects (Biederman, Mezzanotte & Rabinowitz, 1982; Navon, 1977) or may be described more formally as low-level image features corresponding to spatially localized second-order image statistics (Oliva & Torralba, 2001; Torralba & Oliva, 2002, 2003).

Currently, there is not enough evidence to accept either view. The object-first view cannot explain how human observers can recognize scenes even under impoverished viewing conditions such as low spatial resolution (e.g. Schyns & Oliva, 1994) or high clutter (Bravo & Farid, 2006). In such images, object identity information is so degraded that it cannot be recovered locally, yet the scene may still be recognized. Furthermore, research using change blindness paradigms have shown that observers are relatively insensitive in detecting changes to local objects and regions in a scene (Henderson & Hollingworth, 2003; Rensink, O'Reagan & Clark,, 1997; Simons, 2000), suggesting that not all scene objects are represented at once.

While the object-first view cannot explain several key findings in scene perception, the biggest problem for the global view is that it lacks a clear operational definition. Seminal work on artificial stimuli has shown that visual perception tends to proceed in a global-to-local manner (Navon, 1977), but for stimuli as complex as a natural scene, it is not obvious what the global level might be. As Navon (1977) stated, "I am afraid that clear… operational measures for globality will have to patiently await the time that we have a better idea of how a scene is decomposed into perceptual units". Likewise, other authors have noted the need for a grammar of global, scene-emergent properties (Biederman, 1981; Chun, 2003).

In this thesis, I show evidence for a global property representation model view using a variety of behavioral experimental techniques. I propose a grammar of global scene properties to describe the variation in natural landscape scene categories. Some of these global scene properties describe the structure and spatial layout of an environment, such as the degree of *openness*, the degree of *perspective* or the *mean depth* (or volume) of the space (Oliva & Torralba, 2001). Other global scene properties reflect actions that an agent could take in a given environment, such as how well one could *navigate*, or whether one could be *concealed* in the environment, (e.g. affordances, Gibson, 1979). Last, some global scene properties describe the constancy of the scene's surfaces, or how fast they are changing in time. *Transience* is a global scene property depicting the rate at which scene surface changes occur, or alternatively stated, the probability of surface change from one glance to the next. On the other hand, *temperature* describes the differences in visual appearance of a place during the changes of daytime and season, ranging from the intense daytime heat of a desert, to a frigid winter mountain. Using this approach, a forest scene would be described as a *natural*, *enclosed* place with high potential for *concealment* and moderate *temperature* instead of as a collection of trees and leaves.

These properties are studied here as a proof of concept to demonstrate that rapid basic-level scene classification might be mediated through an initial representation that contains global information of scene structure, constancy and function, and not necessarily object information.

In Chapter 2, four experiments explore the human sensitivity to these global properties for rapid scene categorization, as well as examine the computational sufficiency of global properties for basic-level scene classification, comparing the results of a global property based scene representation to a representation built from a scene's objects. In this work, I show that human observers are sensitive to global property information, and that similarity in a global-property space predicts the false alarms made by observers in a rapid basic-level categorization task, whereas object-based models failed to reproduce human errors.

Chapter 3 examines the time course of global property perception relative to the perception of a scene's basic-level category. If the initial representation of a scene contains substantial global property information that allows subsequent basic-level categorization, then observers should require less image exposure to correctly classify a scene's global property than to categorize it at the basic level. This prediction is tested through examining the image presentation time necessary to achieve equal performance across a variety of global property and basic-level category classifications. Results show that although human observers are remarkably efficient in all classifications, global property classifications could be performed with less image exposure than basic-level category classifications.

Another prediction of a global-property representation model is that the visual system should be continuously updated to structural and functional regularities that are useful for scene recognition and therefore prone to adaptation along these dimensions. Chapter 4 tests this prediction in four experiments probing the existence and nature of global scene property aftereffects. Using a novel rapid serial visual presentation paradigm, aftereffects were observed to several global scene properties (magnitude 6% to 22%). Then, using adaptation to probe for a causal link between the perception of global properties and subsequent basic-level scene categorization, I show systematic alterations in observers' basic-level scene categorization following adaptation to a global property.

Finally, Chapter 5 summarizes the experimental work, exploring the implications and limitations of this approach and detailing additional future experimental work. This thesis provides a proposal grammar for categories of environmental scenes, allowing the generation of testable predictions about a global scene representation framework. This thesis has examined some of these predictions, providing the first behavioral evidence for a global initial scene representation, and showing that we may indeed be able to see the forest without necessarily representing the trees.

References

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.) *Perceptual Organization*. pp. 213-263. Hillsdale, New Jersey: Lawrence Erlbaum.

Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143-177.

Biederman, I., Blickle, T. W., Teitelbaum, R. C., Klatsky, G. J., & Mezzanotte, R. J. (1988). Object identification in nonscene displays. *Journal of Experimental Psychology: Human Learning, Memory, and Cognition*, 14, 456-467.

Bravo, M.J. and Farid, H. (2006). Object recognition in dense clutter. *Perception & Psychophysics*, 68(6):911-918.

Chun, M. (2003) Scene perception and memory. In D. Irwin & B. Ross (Eds) *Cognitive Vision.* Elsevier Science & Technology.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory of scene gist. *Journal of Experimental Psychology: General*, 108, 316-355.

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

Henderson, J. & Hollingworth, A. (2003). Global transsaccadic change blindness during scene perception. *Psychological Science* 14(5), 493-497.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 604-610.

Joubert, O., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. *Vision Research*, 47: 3286-3297.

Kaplan, S. (1992). Environmental Preference in a Knowledge-Seeking, Knowledge-Using Organism. In J. H. Barkow, L. Cosmides, and J. Tooby (Eds.) *The Adaptive Mind.* New York: Oxford University Press, 535-552.

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

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42, 145-175.

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

Rayner, K. (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.

Rensink, R. A. O'Regan, J. K. Clark, J. J. (1997). To See or Not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5), 367-373.

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

Simons, D. (2000). Current approaches to change blindness. *Visual Cognition*, 7(1), 1-15.

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

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Pattern Analysis and. Machine Intelligence*, 24, 1226-1238.

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

Van Rullen, R., & Thorpe, S. (2001). The time course of visual processing: from early perception to decision making. *Journal of Cognitive Neuroscience*, 13(4), 454-461.

Chapter 2

Recognition of Global Properties from Global Properties

Published as: Greene, M.R., & Oliva, A. (2009) Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2), 137-176.

1 – Introduction

One of the greatest mysteries of vision is the remarkable ability of the human brain to understand novel scenes, places and events rapidly and effortlessly (Biederman, 1972; Potter, 1975; Thorpe, Fize & Marlot, 1996). Given the ease with which we do this, a central issue in visual cognition is determining the nature of the representation that allows this rapid recognition to take place. Here, we provide the first behavioral evidence that rapid recognition of real-world natural scenes can be predicted from a collection of holistic descriptors of scene structure and function (such as its degree of openness or its potential for navigation), and suggests the possibility that the initial scene representation can be based on such global properties, and not necessarily the objects it contains.

1.1 – Rapid basic-level scene categorization

Human observers are able to understand the meaning of a novel image if given only a single fixation (Potter, 1975). During the course of this glance, we perceive and infer a rich collection of information, from surface qualities such as color and texture (Oliva & Schyns, 2000; Rousselet , Joubert & Fabre-Thorpe, 2005); objects (Biederman, Mezzanotte & Rabinowitz, 1982; Fei-Fei , Iyer, Koch & Perona, 2007; Friedman, 1979; Rensink, 2000, Wolfe, 1998), and spatial layout (Biederman, Rabinowitz, Glass & Stacy, 1974; Oliva & Torralba, 2001; Sanocki, 2003; Schyns & Oliva, 1994), to functional and conceptual properties of scene space and volume (e.g. wayfinding, Greene & Oliva, 2006; Kaplan, 1992; emotional valence, Maljkovic & Martini, 2005).

Indeed, from a short conceptual scene description such as “birthday party,” observers are able to detect the presence of an image matching that description when it is embedded in a rapid serial visual presentation (RSVP) stream and viewed for ~100 milliseconds (Potter, 1975; Potter, Staub & O’Connor, 2004). This short description is also known as the basic-level category for a visual scene (Rosch, 1978; Tversky & Hemenway, 1983), and refers to the most common label used to describe a place.

The seminal categorization work of Eleanor Rosch and colleagues has shown that human observers prefer to use the basic-level to describe objects, and exhibit shorter reaction times to name objects at the basic-level rather than at subordinant or superordinant (Rosch, 1978). It is hypothesized that the basic-level of categorization is privileged because it maximizes both within-category similarity and between-category variance (Gosselin & Schyns, 2001; Rosch, 1978). In the domain of visual scenes, members of the same basic-level category tend to have similar spatial structures and afford similar motor actions (Tversky & Hemenway, 1983). For instance, most typical environments categorized as “forests” will represent enclosed places where the observer is surrounded by trees and other foliage. An image of the same place from very close up might be called “bark” or “moss”, and from very far away might be called “mountain” or “countryside”. Furthermore, the characteristic spatial layout of a scene constrains the actions that can be taken in the space (Gibson, 1979; Tversky & Hemenway, 1983). A “forest” affords a limited amount of walking, while a “countryside” might afford more options for navigation because the space is open. Although such functional and structural properties are inherent to scene meaning, their role in scene recognition has not yet been addressed.

1.2 – The object-centered approach to high-level visual recognition

Many influential models of high-level visual recognition are object-centered, treating objects and parts as the atoms of scene analysis (Biederman, 1987; Biederman, Blickle, Teitelbaum, Klatcky & Mezzanotte, 1988; Bulthoff et al., 1995; Fergus, Perona & Zisserman, 2003; Marr, 1982; Pylyshyn, 1999; Riesenhuber & Poggio, 1999; Ullman, 1999). In this view, the meaning of a real-world scene emerges from the identities of a set of objects contained within it, learned through the experience of object co-occurrence and spatial arrangement (Biederman, 1981; Biederman, 1987; De Graef, Christaens & d'Ydewalle, 1990; Friedman, 1979). Alternatively, the identification of one or more prominent objects may be sufficient to activate a schema of the scene, and thus facilitate recognition (Biederman, 1981; Friedman, 1979).

Although the object-centered approach has been the keystone of formal and computational approaches to scene understanding for the past 30 years, research in visual cognition has posed challenges to this view, particularly when it comes to explaining the early stages of visual processing and our ability to recognize novel scenes in a single glance. Under impoverished viewing conditions such as low spatial resolution (Oliva & Schyns, 1997, 2000; Schyns & Oliva, 1994; Torralba, Fergus & Freeman, 2007); or when only sparse contours are kept, (Biederman, 1981; Biederman et al, 1982; De Graef et al, 1990; Friedman, 1979; Hollingworth & Henderson, 1998; Palmer, 1975) human observers are still able to recognize a scene's basic-level category. With these stimuli, object identity information is so degraded that it cannot be recovered locally. These results suggest that scene identity information may be obtained before a more detailed analysis of the objects is complete. Furthermore, research using change blindness paradigms demonstrates that observers are relatively insensitive to detecting changes to local objects and regions in a scene under conditions where the meaning of the scene remains constant (Henderson & Hollingworth, 2003; Rensink, O'Reagan & Clark,, 1997; Simons, 2000). Last, it is not yet known whether objects that can be identified in a briefly presented scene are perceived, or inferred through the perception of other co-occurring visual information such as low-level features (Oliva & Torralba, 2001), topological invariants (Chen, 2005) or texture information (Walker-Renninger & Malik, 2004).

1.3 – A scene-centered approach to high-level visual recognition

An alternative account of scene analysis is a scene-centered approach that treats the entire scene as the atom of high-level recognition. Within this framework, the initial visual representation constructed by the visual system is at the level of the whole scene and not segmented objects, treating each scene as if it has a unique shape (Oliva & Torralba, 2001). Instead of local geometric and part-based visual primitives, this framework posits that global properties reflecting scene structure, layout and function could act as primitives for scene categorization.

Formal work (Oliva & Torralba, 2001, 2002; Torralba & Oliva, 2002, 2003) has shown that scenes that share the same basic-level category membership tend to have a similar spatial layout. For example, a corridor is a long, narrow space with a great deal of perspective while a forest is a place with dense texture throughout. Recent modeling work has shown success in identifying complex real-world scenes at both superordinant and basic-levels from relatively low-level features (such as orientation, texture and color), or more complex spatial layout properties such as texture, mean depth and perspective, without the need for first identifying component objects (Fei Fei & Perona, 2005; Oliva & Torralba, 2001, 2002, 2006; Torralba & Oliva, 2002, 2003; Vogel & Schiele, 2007; Walker-Renninger & Malik, 2004). However, the extent to which human observers use such global features in recognizing scenes is not yet known.

A scene-centered approach involves both global and holistic processing. Processing is global if it builds a representation that is sensitive to the overall layout and structure of a visual scene (Kimchi, 1992; Navon, 1977). The influential global precedence effect (Navon, 1977, see Kimchi, 1992 for a review) showed that observers were more sensitive to the global shape of hierarchical letter stimuli than their component letters. Interestingly, the global precedence effect is particularly strong for stimuli consisting of many-element patterns, (Kimchi, 1998) as is the case in most real-world scenes. A consequence of global processing is the ability to rapidly and accurately extract simple statistics, or summary information, from displays. For example, the mean size of elements in a set is accurately and automatically perceived (Ariely, 2001; Chong & Treisman, 2003, 2005), as is the average orientation of peripheral elements (Parkes, Lund, Angelucci, Solomon & Morgan, 2001); some contrast texture descriptors (Chubb, Nam, Bindman, & Sperling, 2007) as well as the center of mass of a group of objects (Alvarez & Oliva, 2008). Global representations may also be implicitly learned, as observers are able to implicitly use learned global layouts to facilitate visual search (Chun & Jiang, 1998; Torralba et al, 2006).

While all of these results highlight the importance of global structure and relations, an operational definition of globality for the analysis of real world scenes has been missing. Many properties of natural environment could be global and holistic in nature. For example, determining the level of clutter of a room, or perceiving the overall symmetry of the space are holistic decisions in that they cannot be taken from local analysis only, but require relational analysis of multiple regions (Kimchi, 1992).

Object and scene-centered computations are likely to be complementary operations that give rise to the perceived richness of scene identity by the end of a glance (~ 200-300 msec). Clearly, as objects are often the entities that are acted on within the scene, their identities are central to scene understanding. However, some studies have indicated that the processing of local object information may require more image exposure (Gordon, 2004) than that needed to identify the scene category (Potter, 1975; Schyns & Oliva, 1994; Oliva & Schyns, 2000). In the present study, we examine the extent to which a global scene-centered approach can explain and predict the early stage of human rapid scene categorization performance. Beyond the principle of recognizing the "forest before the trees" (Navon, 1977), this work seeks to operationalize the notion of "globality" for rapid scene categorization, and to provide a novel account of how human observers could identify the place as a "forest", without first having to recognize the "trees".

1.4 – Global properties as scene primitives

We propose a set of global properties that tap into different semantic levels of global scene description. Loosely following Gibson (1979), important descriptors of natural environments come from the scene's surface structures and the change of these structures with time (or constancy). These aspects directly govern the possible actions, or affordances of the place. The global properties were therefore chosen to capture information from these three levels of scene surface description, namely structure, constancy and function.

A total of seven properties were chosen for the current study to reflect aspects of scene structure (mean depth, openness and expansion), scene constancy (transience and temperature), and scene function (concealment and navigability). A full description of each property is found in Table 1. These properties were chosen on the basis of literature review (see below) and a pilot scene description study (see Appendix 8.1) with the requirement that they reflect as much variation in natural landscape categories as possible while tapping into different levels of scene description in terms of structure, constancy and function. Critically, the set of global properties listed here is not meant to be exhaustive [1], as other properties such as *naturalness* or *roughness* (the grain of texture and number and variety of surfaces in the scene) have been shown to be important descriptors of scene content (Oliva & Torralba, 2001). Rather, the goal here is to capture some of the variance in how real world scenes vary in structure, constancy and function, and to test the extent to which this information is involved in the representation of natural scenes.

[1] See Appendix 8.2 for a description of the space of global properties.

Table 1:

Structural Properties

Openness [1,2,3,4] represents the magnitude of spatial enclosure. At one pole, there is a clear horizon and no occluders. At the other pole, the scene is enclosed and bound by surfaces, textures and objects. Openness decreases when the number of boundary elements increases.

Expansion [1] refers to the degree of linear perspective in the scene. It ranges from a flat view on a surface to an environment with strong parallel lines converging on a vanishing point.

Mean depth [1,3] corresponds to the scale or size of the space, ranging from a close-up view on single surfaces or objects to panoramic scenes.

Constancy Properties

Temperature [2,4] refers to the physical temperature of the environment if the observer was immersed in the scene. In other words, it refers to how hot or cold an observer would feel inside the depicted place.

Transience [4,5,7] refers to the rate at which the environment depicted in the image is changing. This can be related to physical movement, such as running water or rustling leaves. It can also refer to the transience of the scene itself (fog is lifting, sun is setting). At one extreme, the scene identity is changing only in geological time, and at the other, the identity depends on the photograph being taken at that exact moment.

Functional Properties

Concealment [4,6] refers to how efficiently and completely a human would be able to hide in a space, or the probability of hidden elements in the scene that would be difficult to search for. It ranges from complete exposure in a sparse space to complete concealment due to dense and variable surfaces and objects.

Navigability [2,4,5] corresponds to the ease of self-propelled movement through the scene. This ranges from complete impenetrability of the space due to clutter, obstacles or treacherous conditions to free movement in any direction without obstacle.

Table 1: Description of the seven global properties of natural scenes used in Experiments 1, 2 and 3. The numbers refer to additional references describing the properties ([1] Oliva & Torralba, 2001; [2] Gibson, 1979; [3] Torralba & Oliva, 2002; [4] Greene & Oliva, 2006; [5] Kaplan, 1992; [6] Appelton, 1975).

1.41: Properties of scene structure

Previous computational work has shown that basic-level natural scene categories tend to have a particular spatial structure (or spatial envelope) that is well-captured in the properties of *mean depth, openness* and *expansion* (Oliva & Torralba, 2001; Torralba & Oliva, 2002). In brief, the global property of *mean depth* corresponds to the scale or size of the space the scene subtends, ranging from a close-up view to panoramic environment. The degree of *openness* represents the magnitude of spatial enclosure whereas the degree of *expansion* refers to the perspective of the spatial layout of the scene. Images with similar magnitudes along these properties tend to belong to the same basic-level category: for example, a “path through a forest” scene may be represented using these properties as “an enclosed environment with moderate depth and considerable perspective”. Furthermore, these spatial properties may be computed directly from the image using relatively low-level image features (Oliva & Torralba, 2001).

1.42: Properties of scene constancy

The degree of scene constancy is an essential attribute of natural surfaces (Cutting, 2002; Gibson, 1979). Global properties of constancy describe how much and how fast the scene surfaces are changing with time. Here, we evaluated the role of two properties of scene constancy: *transience* and *temperature*.

Transience describes the rate at which scene surface changes occur, or alternatively stated, the probability of surface change from one glance to the next. Places with the highest transience would show actual movement such as a storm, or a rushing waterfall. The lowest transience places would change only in geologic time, such as a barren cliff. Although the perception of transience would be more naturalistically studied in a movie rather than a static image, humans can easily detect implied motion from static images (Cutting, 2002; Freyd, 1983), and indeed this implied motion activates the same brain regions as continuous motion (Kourtzi & Kanwisher, 2000). *Temperature* reflects the differences in visual appearance of a place during the changes of daytime and season, ranging from the intense daytime heat of a desert, to a frigid snowy mountain.

1.43: Properties of scene function

The structure of scene surfaces and their change over time governs the sorts of actions that a person can execute in an environment (Gibson, 1979). The global properties of *navigability* and *concealment* directly measure two types of human-environment interactions deemed to be important to natural scene perception from previous work (Appelton, 1975; Gibson, 1958, 1979; Kaplan, 1992; Warren, Kay, Zosh, Duchon & Sahuc, 2001). Insofar as human perception evolved for goal-directed action in the environment, the rapid visual estimation of possible safe paths through an environment was critical to survival (Gibson, 1958). Likewise, being able to guide search for items camouflaged by the environment (Merilaita, 2003), or to be able to be concealed oneself in the environment (Ramachandran, Tyler, Gregory, Rogers-Ramachandran, Duessing, Pillsbury & Ramachandran, 1996) have high survival value.

1.5: Research questions

The goal of the present study is to evaluate the extent to which a global scene-centered representation is predictive of human performance in rapid natural scene categorization. In particular, we sought to investigate the following questions: (1) are global properties utilized by human observers to perform rapid basic-level scene categorization? (2) Is the information from global properties sufficient for the basic-level categorization of natural scenes? (3) How does the predictive power of a global property representation compare to an object-centered one?

In a series of four behavioral and modeling experiments, we test the hypothesis that rapid human basic-level scene categorization can be built from the conjunctive detection of global properties. After obtaining normative ranking data on seven global properties for a large database of natural images (Experiment 1), we test the use of this global information by humans for rapid scene categorization (Experiment 2). Then, using a classifier (Experiment 3), we show that global properties are computationally sufficient to predict human performance in rapid scene categorization. Importantly, we show that the nature of the false alarms made by the classifier when categorizing novel natural scenes is statistically indistinguishable from human false alarms, and that both human observers and the classifier perform similarly under conditions of limited global property information. Critically, in Experiment 4 we compare the global property classifier to two models trained on a local region-based scene representation and observed that the global property classifier has a better fidelity in representing the patterns of performance made by human observers in a rapid categorization task.

Although strict causality between global properties and basic-level scene categorization cannot be provided here, the predictive power of the global property information and the convergence of many separate analyses with both human observers and models support the hypothesis that an initial scene representation may contain considerable global information of scene structure, constancy and function.

2 – General method

2.1 – Observers

Observers in all experiments were 18-35 years old, with normal or corrected-to-normal vision. All gave informed consent and were given monetary compensation of $10/hour.

2.2 – Materials

Eight basic-level categories of scenes were chosen to represent a variety of common natural outdoor environments: desert, field, forest, lake, mountain, ocean, river and waterfall. The authors amassed a database of exemplars in these categories from a larger laboratory database of ~22,000 (256x256 pixel) full-color photographs collected from the web, commercial databases, personal digital images and scanned from books (Oliva & Torralba, 2001, 2006). From this large database, we selected 500 images [2] chosen to reflect natural environmental variability. To estimate the typicality of each image, independent, naïve observers ranked each of the 500 images on its prototypically for each scene category, using a 1-5 scale (see Appendix 8.3 for a description of the typicality norming task). The most prototypical 25 images for each of the eight basic-level category were kept, for a grand total of 200 images which were used in Experiment 1-4 (see details in Appendix 8.3). The remaining 300 poly-categorical images were used in Experiment 3, section 5.27. For human psychophysics experiments, we used Matlab and the Psychophysics Toolbox as presentation software (Brainard, 1997; Pelli, 1997).

3 - Experiment 1: Normative rankings of global properties on natural scenes

First, we obtained normative rankings on the 500 natural scenes along the seven global properties. These normative rankings provide a description of each image and basic-level category in terms of their global structural, constancy and functional properties. Namely, each image is described by 7 components, each component

[2] The image database may be viewed on the authors' web site.

representing the magnitude along each global property dimension (see examples in Figure A2 in Appendix 8.2).

As Experiments 2-3-4 involve scene categorization using global property information, robust rankings are essential for selecting images for the human psychophysics in Experiment 2 as well as for training and testing the classifier used in Experiment 3.

3.1 – Method

3.11 – Participants

55 observers (25 males) ranked the database along at least one global property, and each property was ranked by at least ten observers.

3.12 – Procedure

Images were ranked using a hierarchical grouping procedure (Figure 1, Oliva & Torralba, 2001). This allows the ranking of a large number of images at once, in the context of one another.

For a given global property, each participant ranked two sets of 100 images. The two halves of the database were pre-chosen by the authors to contain roughly equal numbers of images in each semantic category. 100 picture thumbnails appeared on an Apple 30" monitor (size of 1.5 x 1.5 deg / thumbnail). The interface allowed participants to drag and drop images around the screen and to view a larger version of the image by double-clicking on the thumbnail.

Participants were given the name and description of a global property at the start of a ranking trial. They were instructed to divide the images into two groups based on a specific global property such that images with a high magnitude along the global property were placed on the right-hand side of the screen while images with a low magnitude were placed on the left (see Figure 1). In a second step, participants were asked to split each of the two groups into two finer divisions, creating four groups of images that range in magnitudes along the specified global property. Finally, the four groups were split again to form a total of 8 groups, ordered from the lowest to the highest 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. Participants repeated this hierarchical sorting process on the remaining 100 pictures in database along the specified global property. Participants had unlimited time to complete the task, but on average completed a trial in 30 minutes. As the task was time consuming, not all participants ranked all seven properties, and we are reporting results from 10 observers per property, normalized to fit in the range of 0 to 1.

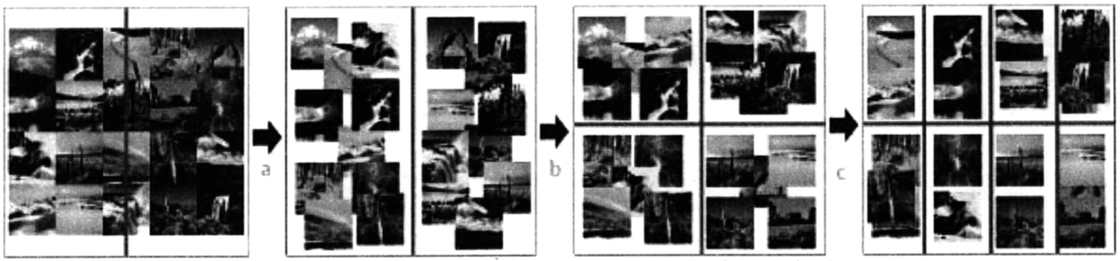


Figure 1: A schematic illustration of the hierarchical grouping task of Experiment 1. Here, a ranking along the global property *temperature* is portrayed. a) the images are divided into two groups with the "colder" scenes on the left and the "warmer" scenes on the right; b) Finer rankings are created by dividing the two initial groups into two subgroups and c) Images in each quadrant are again divided into two subgroups to create a total of eight groups, ranked from the "coldest" scenes to the "hottest" scenes.

3.2 – Results

3.21 – General description

Examples of images that were ranked as low, medium and high for each global property are shown in Figure 2. Global properties are continuous perceptual dimensions, and therefore image ranks spanned the range of possible values across the database (Scattergrams of rankings by category for all global properties can be seen in Figure A1, Appendix 8.2). It is essential to note in Figure A1 the high scatter of rankings indicates that the basic-level category label is not the determinant of the global property ranking for any single global property. In other words, *concealment* is not just another way of saying *forestness*.

In order to compare the time-unlimited rankings of Experiment 1 to the speeded categorization task of Experiment 2, it is necessary to know that global properties can be rapidly and accurately perceived by human observers. Furthermore, a similar ranking of images along global properties when presentation time is limited ensures that the rankings of Experiment 1 are not due to inferences based on the scene schema. To this end, we ran a control speeded classification task [3] (see the description of this experiment in Appendix 8.4). Results showed that indeed, global properties could be estimated from limited presentation time. The mean correlation of the speeded classification to the hierarchical rankings was 0.82, ranging from 0.70 for concealment to 0.96 for temperature (all significant), see Appendix 8.4 for more details.

[3] Appendix (8.4) describes a speeded classification task, to verify that the global properties of natural images are perceived under conditions of limited presentation time. The logic, as suggested by an anonymous reviewer, is that under limited presentation time, the perception of global properties might be less contaminated by other semantic information about the scene category. Although category information cannot be completely abolished in a short presentation time, other data in a forthcoming article by the authors show that the detection of global properties in a scene is significantly better than the detection of the same scene's basic-level category at a 20ms presentation time (see also Joubert et al, 2007 showing that the global property of naturalness is available faster than a scene's basic-level category), indicating that some category information was suppressed in the manipulation.

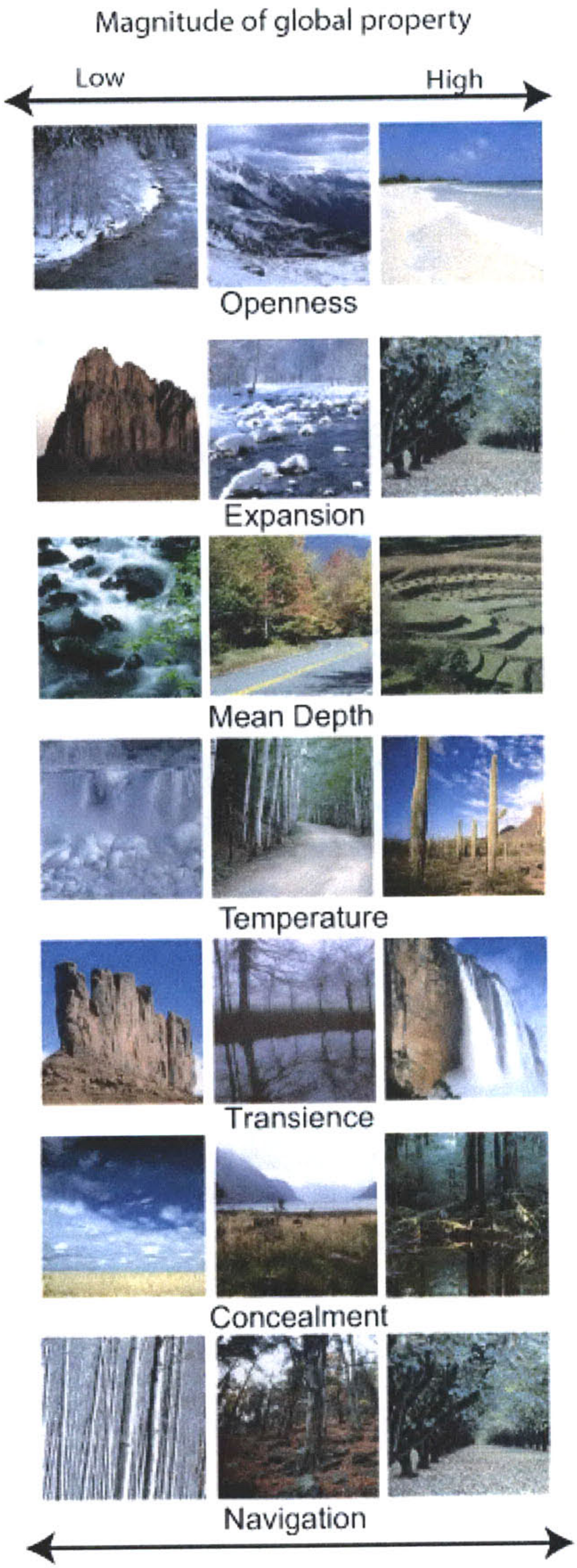


Figure 2: Examples of scenes with low, medium and high rankings from Experiment 1 along each global property.

3.22 – Between-observer consistency in ranking images along each global property

The extent to which global properties represent a reasonable basis for natural scene recognition depends critically on the extent to which the global properties can be ranked consistently by human observers.

Here we are using the 200 prototypical images as they give strong ground truth for the purpose of categorization in Experiments 2-4. We computed observers' consistency as a Spearman's rank-order correlation for each possible pairing of observers for all seven global properties. The mean and standard error for these correlation coefficients by global property are shown in Table 2. Between-observer Spearman's rank-order correlations ranged from 0.61 (transience) to 0.83 (openness), and were all statistically significant ($p < 0.01$). This indicates that different observers estimated the degree of these global properties in similar ways (see also Oliva & Torralba, 2001; Vogel & Schiele, 2007 for similar results) and agreed well on which images represented a high, medium and low magnitude for a given global property.

	Openness	Expansion	Mean depth	Temperature	Transience	Concealment	Navigability
r	0.83	0.64	0.76	0.73	0.61	0.65	0.69
s.e.m	0.01	0.01	0.01	0.01	0.01	0.01	0.01

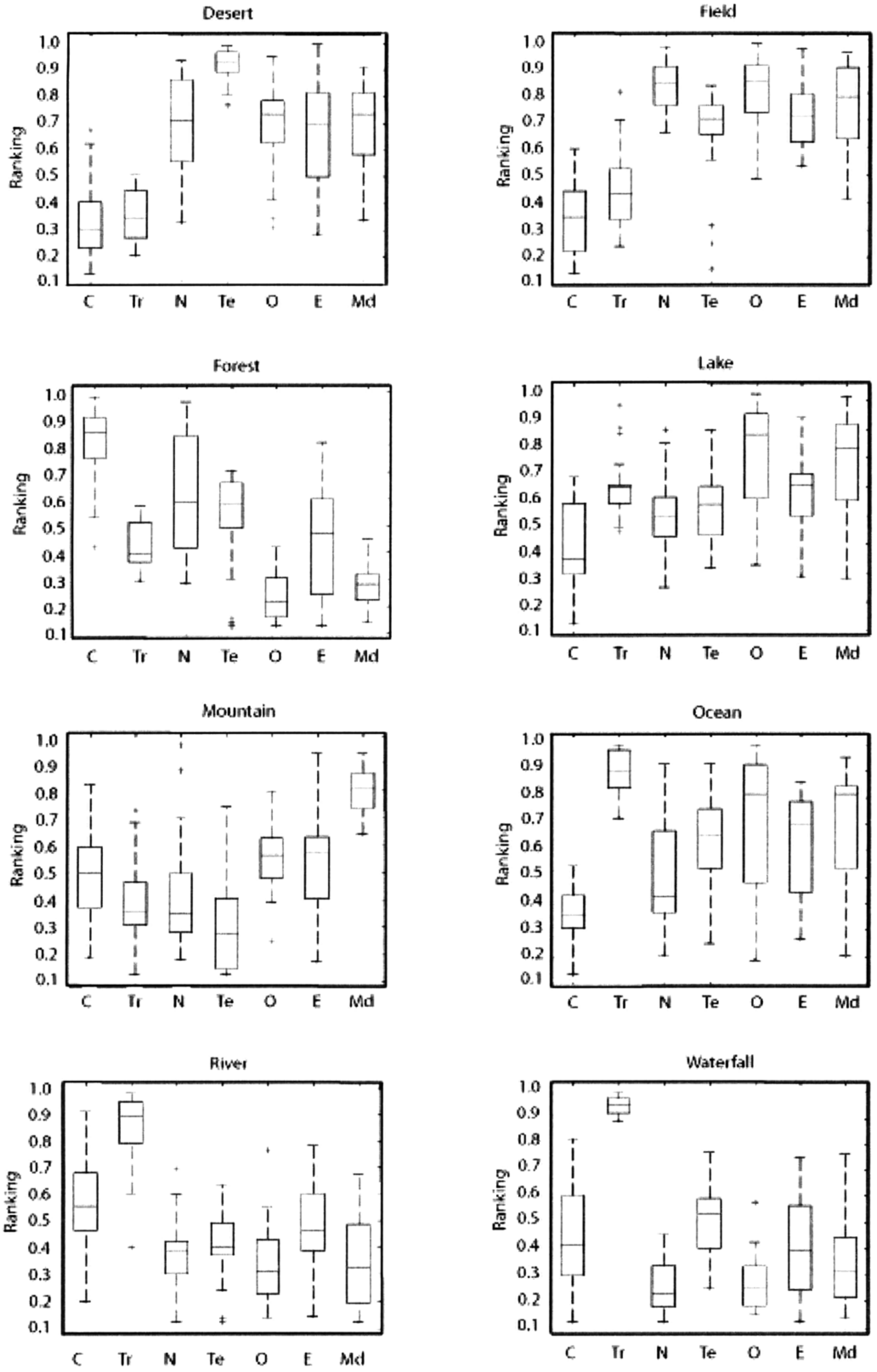
Table 2: Spearman's rank order correlations along with standard error of the mean between observers for each global property from the rankings given in Experiment 1.

3.23 – Global property descriptions of semantic categories

The subsequent experiments test the utility of a global property representation for rapid scene categorization. In this representation, images are represented as points in a seven-dimensional space where each axis corresponds to a global property. How are different basic-level categories described in this space?

To visualize the global property signature of each semantic category, we computed the category means and ranking spread for each global property. Figure 3 shows box-and-whisker plots for the global property rankings for each category, creating a conceptual signature of the category. For example, most deserts were ranked as *hot*, *open* and *highly navigable* environments, with a low magnitude along the *transience* and *concealment* dimensions while most waterfalls are *closed*, *highly transient* environments that are *less navigable*. Other categories, such as lakes, have global property ranking averages that were intermediate along each dimension, meaning that most lakes have a medium level of *openness* and *expansion*, are neither environments perceived as *very cold* or *very warm*, and so on.

Euclidean distance measures between each pair of basic-level categories provided a *conceptual distance* metric between basic-level categories (see Table A4 and details in Appendix 8.6). As expected from intuition, categories like *waterfall* and *river* are close to each other, but categories like *field* and *waterfall* are very distant.


Desert

Field

Forest

Lake

Mountain

Ocean

River

Waterfall

Figure 3: Box-and-whisker plots of global property rankings for each semantic category, calculated from the ranking data in Experiment 1. Properties are, right to left, C=Concealment, Tr=Transience, N=Navigability, Te=Temperature, O=Openness, E=Expansion and Md=Mean depth. Lines indicate

median rankings, boxes indicate quartiles and whiskers indicate range. Significant outlier images are shown as crosses.

3.3 – Discussion

Here we have shown that observers can provide normative rankings on global properties with a high degree of consistency. We have also provided a conceptual description of basic-level category prototypes as the mean global property rankings of a category.

To what extent do the scene-centered semantic category descriptions shown in Figure 3 contribute to human observers' mental representations of scene identity? We test this explicitly in Experiment 2.

4 – Experiment 2: Human use of global properties in a rapid scene categorization task

The goal of Experiment 2 was to test the extent to which global property information in natural scenes is utilized by human observers to perform rapid basic-level scene categorization. A global property-based scene representation makes the prediction that scenes from different semantic categories but with similar rankings along a global property (e.g. oceans and fields are both *open* environments) will be more often confused with each other in a rapid categorization task than scenes that are not similar along a global property (e.g. an *open* ocean view and a *closed* waterfall). We tested this hypothesis systematically by recording the false alarm rates for each basic-level category (serving as targets in blocked yes-no forced choice task) when viewed among distractor images that all shared a particular global property pole (such as *high concealment* or *low openness*).

4.1 – Method

4.11 – Participants

For a purpose of completeness and replication of our effects, two groups of participants participated in Experiment 2. First, four participants (1 male) completed the entire experimental design. Throughout the experiment, we will refer to this group as the *complete-observer* group. While having all observers complete all blocks of the experiment is statistically more robust, it could also lead to over learning of the target images. To eliminate the learning effect, a *meta-observer* group consisting of 73 individuals (41 male) completed at least 8 blocks (400 trials) of the design, for a total of eight meta-observers (see Appendix 8.5 for details on the analysis of meta-observer data). Meta-observer analysis is justified here because the critical analyses are on the image items.

4.12 – Design

The experimental design consisted of a full matrix of target-distractor blocks where each basic-level category was to be detected amongst distractor images from different semantic categories that shared a similar ranking along one global property. Both high and low magnitudes of each global property were used, yielding 112 blocked conditions (8 target categories x 7 global properties x 2 magnitudes). For example, a block would consist of one semantic category (such as forest) seen among images that were all ranked in Experiment 1 as (for example) high-transience. The distractor sets were chosen to reflect a wide variety of semantic categories, and to vary in other global properties while keeping ranks in the manipulated property constant. Therefore, as best as possible, global properties were independently manipulated in this design. Distractor sets for a given global property magnitude were therefore chosen uniquely for each category. High and low rankings were defined as imaged ranked as >0.6 and <0.3 for a given global property.

4.13 – Procedure

Each of the 112 experimental blocks contained 25 target and 25 distractor images. At the start of each block, participants were given the name of the target category and were instructed to respond as quickly and accurately as possible with a key press ("1" for yes, "0" for no) as to whether each image belonged to the target category. Each trial started with a 250 msec fixation cross followed by an image displayed for 30 msec, immediately followed by a 1/f noise mask presented for 80msec. Visual feedback (the word "error") followed each incorrect trial for 300msec.

4.2 – Results

For all analyses, we report results for both the complete-observer and the meta-observer groups. Results from the two groups support each other well. In addition to providing a self-replication, examining individuals completing the entire design reduces the noise seen from pooling individual performances. On the other hand, the meta-observer group reduces the problem of over-learning the target images.

In the following, we report 4 different analyses on both correct detection (hit) and false alarms: 4.21 – the general performance of human observers in rapid basic-level scene categorization; 4.22 –the power of target-distractor global property resemblance in predicting which particular images will yield false alarms to a basic-level category target.; 4.23 - the relation between false alarms made between basic-level categories and the relative distances of those categories in global property space; 4.24 – the effect of global property similarity on reaction time.

4.21 – Basic-level scene categorization: overall performance

The complete-observers' average hit rate was 0.87 with a mean false alarm rate of 0.19. This level of performance corresponds to an average *d'* sensitivity of 2.07. Performance by semantic category is detailed in Table 3. With this short 30 millisecond presentation time, observers could reliably detect all scene categories (all *d'*>1.0). However, critical for subsequent analyses, observers made substantial false alarms to each category as well, giving a rich range of performance data to work with.

For the 8 meta-observers, the mean hit rate was 0.78, with a mean false alarm rate of 0.24. This corresponds to a *d'* of 1.58. For the complete-observer group, we looked at hit rate across the 14 times they viewed the target images. For each observer, we performed a linear regression on the hit rate over these blocks and found that for 3 of the 4 subjects, there was a positive slope (mean – 0.095, just under 1% per block), indicating that there was learning of the targets over the course of the experiment.

	Hit	False alarm	d'
Desert	0.83 (0.88)	0.18 (0.17)	1.88 (2.13)
Field	0.77 (0.88)	0.30 (0.20)	1.27 (2.02)
Forest	0.88 (0.96)	0.17 (0.11)	2.23 (2.97)
Lake	0.74 (0.91)	0.26 (0.18)	1.32 (2.28)
Mountain	0.78 (0.88)	0.25 (0.17)	1.50 (2.15)
Ocean	0.68 (0.87)	0.27 (0.25)	1.11 (1.79)
River	0.69 (0.89)	0.30 (0.23)	1.03 (1.97)
Waterfall	0.91 (0.95)	0.20 (0.16)	2.29 (2.67)

Table 3: Overall human performance in rapid categorization task of Experiment 2. Shown are hit rate, false alarm rate and sensitivity measure d', measured as the mean for each category over eight meta-observers. Numbers in parentheses show the same measurements for the complete-observer design.

4.22 – The role of global properties on basic-level categorization performance

A prediction of the scene-centered approach is that distractor images that share a global property ranking with the target prototype should yield more false alarms than images that are less similar to the target prototype. A pictorial representation of sample results is shown in Figure 4: forests, which tend to be closed (c.f. Figure 3) have more false alarms to closed distractors than to open distractors, and the opposite is true of fields, which tend to be open environments.

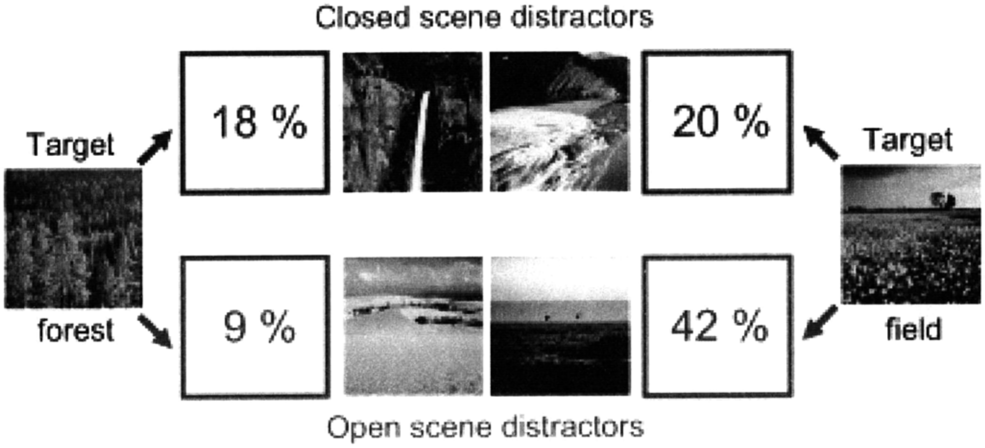


Figure 4: Illustration of human performance along different distractor sets in Experiment 2. Distractor sets that share a global property with the target category (*closed* is a property of forests and *open* is a property of fields) yield more false alarms than distractor sets that do not. Representative numbers taken from meta-observers' data.

A global property-based scene representation would predict that any image's confusability to any target category could be predicted from this image's global property distance to the target category. For example, in general, mountain scenes were ranked as moderately-low navigability (c.f. Figure 3). Therefore, in a block where mountains were to be detected among low-navigability distractors, we would expect more false alarms to distractors that are also moderately-low navigability than non-navigable distractors of greater magnitude (such as a very dense forest).

In each of the 112 experimental blocks, a single semantic category was to be detected among distractor images that had a common global property rank. For each distractor image in these blocks, we computed the one-dimensional distance between its global property rank on the manipulated global property to the mean global property rank of the target category for the same property. For example, in a block where deserts were viewed among low-expansion scenes, each distractor would be expressed as the distance between its rank on expansion (given from Experiment 1), and the mean desert rank for expansion (c.f. Figure 3).

Therefore, all of the distractor images in the entire experiment could be ranked from most similar to the target category to least. If global property information is used to help human observers estimate the image category, then global property resemblance should predict the false alarms that are made during the experiment.

To test, we first binned the false alarm data into quartiles based on ascending target-distractor distance. The mean percent correct rejections for each quartile for each data set are shown in Table 4. For both groups, the accuracies increase monotonically with distance, indicating that difficulty of image categorization is in part due to the resemblance of the distractors to the target category prototype. Human categorization performance is not obliterated by this one-dimensional similarity, however as even the most similar 1% of distractors are still classified significantly above chance by the meta-observers: 64% correct, $t(198)=5.5$, $p<0.0001$.

Quartile	25	50	75	100
% Correct Rejection (meta-observers)	71.9	74.5	78.1	82.1
% Correct Rejection (complete-observer)	75.9	78.1	84.4	88.5

Table 4: Average human correct rejection performance for both experimental groups in Experiment 2 on distractor images arranged from smallest distance to target category prototype to largest. Performance suffers with decreasing distance to target prototype, but remains above chance.

We also performed a correlation on the distractor distance data, using the mean false alarm rate for each distractor to its distance from target prototype mean. For the complete-observer group, we found a striking relation with correlation coefficients ranging from 0.98 to 0.91, when binning the data respectively in 8 bins and 25 bins (for all correlations, $p<0.0001$). For the meta-observers, correlations ranged from 0.95 for 8 bins, to 0.81 for 25 bins, all correlations were significant ($p<0.001$).

This strong relation shows that images that resemble the category global property prototype are more often mistaken with the target category than other images, and suggests that with a short presentation time, global property information is used by human observers to categorize natural scenes into basic-level categories.

4.23 – Distance in global property space predicts pairwise category false alarms

Are some semantic categories confused with each other more often than others? Can such asymmetries be understood through a scene-centered global-property representation? Ashby & Lee (1991) showed that false alarms increase with increasing similarity between targets and distractors in a detection task. Therefore, if our global properties are incorporated into the human scene representation, we would expect false alarms made between semantic categories in the rapid categorization task to follow from the categories' similarity in global property space (from Experiment1, see Figure 3).

As the experimental task was a yes-no forced choice within a block of uniform target categories, the false alarms made in a given block provide insight into which category observers believed the image to be. For example, a false alarm to a forest image while looking for river targets indicates that the observer believed the picture of the forest to be a river. False alarm rates between each pair of categories were thus computed (see Appendix 8.6 for more details).

We then computed the Euclidean distance between each category in the global property space (n*(n-1)/2 = 28 pairwise comparisons for the n=8 categories). This is a dissimilarity metric: larger values indicate more differences between two categories (See Appendix 8.5 for more details).

For the complete-observer group, we found a strong negative correlation between category dissimilarity and false alarm rates ($r=-0.76$, $p<0.001$), indicating that pairs of categories that are similar in global property space (such as river and waterfall) are more often confused by human observers than pairs of categories that are more distant, such as field and waterfall. The same pattern held for the meta-observers: ($r=-0.78$, $p<0.001$).

4.24 – The reaction time effects of global property similarity

The previous analyses have shown that the probability of human observers mis-categorizing images given a brief presentation is strongly related to how similar a given distractor is to the target category in global property space. Is there evidence of global property similarity for the images that are correctly categorized? In particular, is the speed at which an image can be correctly categorized inversely related to how similar it is to the category prototype? One can imagine that a distractor sharing very few global properties with the target category might be more quickly rejected than a distractor that more closely resembles the target category.

For this analysis, we report data from the complete-observer group as individual differences in reaction time from the meta-observer group are confounded in the blocked design. For all correctly rejected distractors, we correlated the participants' reaction time to the Euclidean distance of that distractor to the target category in global property space. We found that there was a strong inverse relation between target-distractor resemblance and reaction time (r=-0.82, p<0.0001), indicating that distractors that are more dissimilar to the target category are more quickly rejected than distractors that are more similar. In other words, similarity in global property space predicts the mistakes that human observers tend to make as well as which images will take longer to categorize.

4.3 – Discussion

The previous analyses have shown that with a very brief image exposure, human observers are able to detect the basic-level category of a natural scene substantially above chance (section 4.21; see also Joubert, Rousselet, Fize & Fabre-Thorpe, 2007; Oliva & Schyns, 2000; Potter, 1975; Rousselet et al., 2005). However, participants' performances were far below ceiling, suggesting that the scene representation afforded by this amount of image exposure was incomplete, providing a rich array of false alarms that are useful for understanding the initial representation.

In this experiment, we have shown converging evidence from different analyses indicating that human observers are sensitive to global property information during very brief exposures to natural images, and that global information appears to inform basic-level categorization.

First, we have shown that the probability of false alarm to a given image can be very well predicted from the one-dimensional distance of this image's rank along a global property to the target category prototype for that same property (section 4.22). We have also shown that semantic categories that are more often confused by human observers are more similar to one another in global property space (section 4.23). As distractor images varied in semantic categories, other global properties and objects, this implies that global property information makes up a substantial part of the initial scene representation. Last, we have shown that the reaction times for correctly rejected distractors were also related to the distractors' resemblance to the target category (4.24). Altogether, these results support a scene-centered view of scene understanding that asserts that spatial and functional global properties are potential primitives of scene recognition.

5 – Experiment 3: The computational sufficiency of global properties for basic-level scene categorization

We have shown so far that global property information strongly modulates human performance in a rapid scene categorization task. To what extent is a global property representation sufficient to predict human rapid scene categorization performance? To answer this question, we built a conceptual naïve Bayes classifier whose only information about each scene image was from the normative ranking data of Experiment 1. Therefore, the classifier is agnosic to any other visual information (color, texture, objects) that human observers could have used to perform the task. Here we compare the performance of this classifier (correct and false alarms) to the human scene categorization performance of Experiment 2.

5.1 – Method

The training input to the classifier consisted of the ranks that each image received for each of the seven global properties along with a label indicating which semantic category the image belonged to. From this input, the classifier estimated Gaussian distributions for each category along each global property. Then, given a test image (not used in training), the classifier computed the most likely semantic category for the set of global properties given to it:

$$C_{est} = \arg\max_{c \in C} \sum_{k=1}^{k} \ln \frac{1}{\sqrt{2\pi\sigma_{jk}^{2}}} - \frac{1}{2\sigma_{jk}^{2}}(x-\mu_{jk})^{2}$$

where the log likelihood of each category *j* is estimated from the distributions of each property dimension *k* (for background, see Mitchell, 1997). For a discussion on the assumptions of such a classifier, see Appendix 8.6.

The classifier was run 25 times, testing each image in turn using a leave-one-out design. 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.

It is of note that the naïve Bayes classifier was chosen to be the simplest classifier for testing this global property representation. All reported results were also done with a linear discriminant analysis with no significant performance differences (see Appendix 8.7).

5.2 – Results

In comparing a classifier's performance to human performance for the goal of gaining insight into the human representation, it is necessary to examine classifier performance at several levels. Similar overall performance is not enough since any psychophysical task can be made arbitrarily harder or easier by changing presentation time, for example. The errors made by a classifier are more informative than the overall correct performance because similarities in errors make a stronger argument for a similar representation. Conversely, dissimilarities in the patterns of errors are informative in refining hypotheses. We report here four distinct types of analyses using data from Experiments 1, 2 and 3: section 5.21 – the overall performance of the classifier relative to human scene categorization performance from Experiment 2; sections 5.22– an examination of the types of classification errors made by both humans and classifier; section 5.23 – an examination of the distances between categories in our scene-centered space (Experiment 1) and how this predicts errors made by both classifier and human observers; and sections 5.24 and 5.25– a comparison of how the classifier and human observers perform under conditions where a complete global property representation cannot be used for scene categorization. As a last test of this model (5.26), we compare the classifier's responses to non-prototypical images to that of the human norming data of Experiment 1 (see Appendix 8.3).

5.21 – Classifier performance: Percent correct and correlation to human basic-level category performance

Overall, the performance of the classifier was remarkably similar to that of human meta-observers: the overall proportion correct for the classifier was 0.77 (0.77 for human meta observers, t(7)<1). The performance for the complete observer group was higher (proportion correct was 0.86), in part because of the over-learning of the stimuli.

To get an idea of how classifier performance compared to human performance by basic-level category, we correlated meta-observer's correct performance and classifier correct performance and found a striking similarity: the by-category correlation was r=0.88, p<0.01 (see Figure 5). This level of agreement did not differ from meta-observer agreements (r = 0.78: t(7)=1.72, p=0.13), indicating that the classifier's overall correct performance and correct performance by category were indistinguishable from human performances. Similarly, the correlation between the classifier and the mean correct performance of the complete observer group was similarly high (r=0.75, p<0.01).

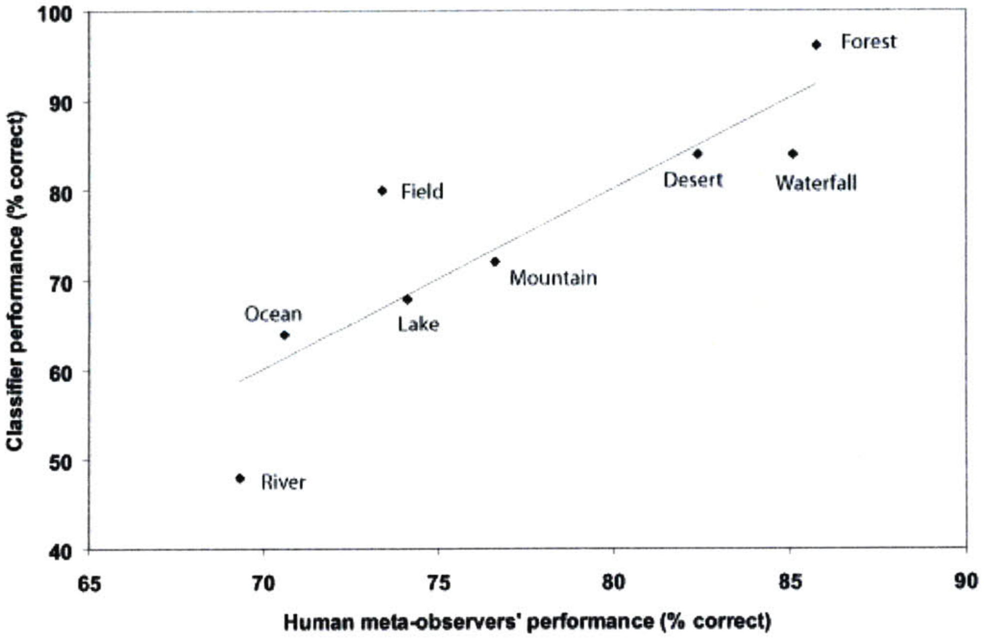


Figure 5: Categorization performance (percent correct) of naïve Bayes classifier in Experiment 3 is well-correlated with human rapid categorization performance from Experiment 2 (meta-observer data).

5.22 – Error analysis: Easy and Difficult Images

Do human observers and the classifier have difficulty classifying the same images? We looked at the errors that both humans and classifier made in a by-image item analysis, comparing the probability of classifier failure (average performance in 4, 10 and 25 bins, due to the binary classification of the 200 images by the classifier) to human false alarm rates (over the same bins).

We found a significant correlation between the classifier and the meta-observers (for 10 bins, $r=0.89$, $p<0.0001$) indicating that indeed humans and classifier have trouble categorizing the same images. Bin size did not affect the nature of the result: using bin sizes of 4 and 25, the correlation coefficients were 0.97 and 0.76 respectively (all significant). Similarly, the correlation between the classifier and participants from the complete-observers design were all significant ($p<0.001$, $r=0.96$, $r=0.81$, and $r=0.64$ for the same bin sizes).

5.23 – Qualitative error analysis: Distribution of error types

Next, we sought to determine the qualitative similarity of the false alarms made by both classifier and human observers. The yes-no forced choice task of the human observers allowed insights into which category observers believed an image to be given a false alarm, and this can be compared directly to the output of the classifier. In other words, in a block where the target image was *river*, and an observer made a false alarm to an image of a forest, does the classifier also call this forest a river?

Given an error made by the classifier, we found that at least one human observer in the meta-observer group made the same false alarm in 87% of the images (88% for the complete-observer group). However, human observers are also prone to finger errors, attentional lapses and other possible mistakes, so when we include only the false alarms that at least five of the eight meta-observers made; there was human-classifier correspondence on 66% of the images (59% for at least 3 of the 4 participants who completed the entire experiment).

Examples of the correct responses and the false alarms made by the classifier and human observers (meta-observer group) are shown in Figure 6. This indicates that the scene categorization performance of a classifier knowing *only* about global property rankings is highly similar to that of human observers when given a 30 msec exposure to a scene image.

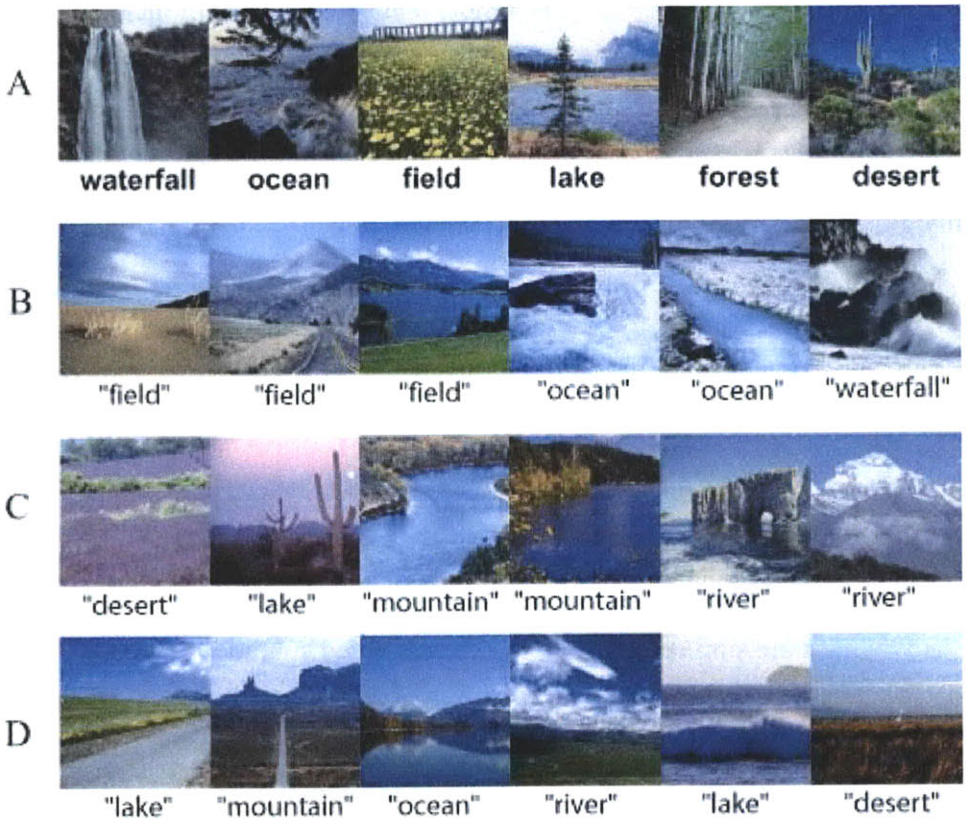


Figure 6: Examples of human and model performances. (A) (bold titles) corresponds to the correct responses made by both humans (Experiment 2) and the global-property classifier (Experiment 3) for the above scene pictures. The other rows (with titles in quotes) represent categorization errors made

respectively by both humans and the model (B); by the model only (C); by the humans only (D), for the respective scene pictures.

We have shown so far that the overall performance of the global property classifier as well as the types of errors it made is highly similar to the performance of human observers in a rapid scene categorization task. To further compare classifier to human performance, we created a category-by-category confusion matrix for the global property classifier (see false alarms Table A6 in Appendix 8.6) and human observers (human matrix of false alarms from Experiment 2, see Table A5 in Appendix 8.6). We found that the between-category confusions made by the classifier were highly correlated with those made by human observers ($r=0.77$, $p<0.0001$ for complete observers and $r=0.73$ for the meta-observers, $p<0.0001$). It is of note that the diagonals of the confusion matrices (the correct detections) were taken out for both as it would have led to a spuriously high correlation. This analysis further suggests that a scene representation containing only global property information predicts rapid human scene categorization, a result which strengthens the hypothesis that a global scene-centered representation may be formed by human observers at the beginning of the glance.

5.24 – "Knocking out" a global property I: missing properties

A stronger case for a global scene representation in human observers would be if the classifier and humans are similarly impaired under degraded conditions. We have shown so far that these global properties are sufficient to predict human performance in rapid scene categorization. From Experiment 2, we found that human observers are remarkably flexible in scene categorization under conditions where target-distractor similarity along a global property dimension decreases the utility of that dimension for categorization - performance suffers but remains above chance with such incomplete information. How does the classifier perform when similarly impaired? To test, we compared human false alarms in Experiment 2 to runs of the classifier trained with all global properties but one in turn. Experiment 2 "knocked-out" global properties for human observers by matching the target and distractors on that property, reducing the utility of the property for categorization. For example, assuming *high transience* is a diagnostic property of oceans, classifying oceans among *high transience* scene distractors will render *transience* useless for the task. Likewise, training the classifier without a property "knocks-out" that property because there is no representation of the property at all.

All training and testing procedures were identical to the previously presented method in section 5.1 except that all images were represented by six global properties instead of the full set of seven, which served as a performance baseline. For the human comparison, for each global property we used the pole (high or low rank) that yielded the most false alarms. For each category, we compared these false alarm rates to the average performance of that category over all distractor conditions.

For each basic-level category we compare the increase in false alarms for the classifier to the increase in false alarms for human observers. Interestingly, "knocking-out" the use of a global property decreased performance to a similar degree: overall increase in false alarms by category was an average of 5.2% more for the classifier and 3.2% more for the complete observer group (3.1% for meta-observers, difference between humans and model were not significant, t(7) < 1) indicating that the loss of global property information affected both human observers and the model to a similar degree, and that the classifier's representation was similarly robust to incomplete global property information. Furthermore, the correlation between classifier and human correct performance by category remains strong in this manipulation ($r=0.81$, $p<0.0001$ for the complete-observers, and $r=0.83$ for meta-observers), indicating that the absence of each global property is similarly disruptive to categorization, and suggesting that both observer types are using similar diagnostic global property information to perform the categorization task. Again, the correlation existing between the classifier and mean human performance was not different from the agreement between meta-observers ($t(7)<1$), indicating that the classifier's performance is indistinguishable from human observers.

5.25 – "Knocking out" a global property II: the role of all properties

What is the limit of the classifier's ability to deal with incomplete information and to what extent are all of the global properties necessary to predict human categorization performance? To address this question, we ran the classifier on exhaustive combinations of incomplete global property data, from one to six global properties.

The average performance of the classifier for each number of global properties used is shown in Figure 7a. Interestingly, when the classifier is trained on only one of the global properties, categorization performance is still significantly above chance (30%, chance being 12.5%, $t(6)=7.93$, $p<0.0001$) and reaches a plateau when combinations of six global properties are used (74%).

Next, we looked at which combinations of global properties lead to maximum performance for all eight basic-level categories. We tabulated the average performance of global property combinations containing each global property. If the maximum classifier performance is carried by one or two properties, one would expect maximum performance when these properties are present and diminished performance with other combinations. Instead, Figure 7b shows that all properties were represented in these combinations with similar frequency (between 54-61% correct). Although global property combinations containing *transience* are slightly higher than the mean performance ($t(6) = 2.0$, $p<0.05$), and combinations containing *expansion* trend toward lower performance ($t(6) = 1.8$, $p=0.12$), this result suggests that overall categorization performance is not carried by one or two global properties, but rather that each global property provides essential information for classifying all 8 basic-level categories. This result is conferred by the multi-dimensional scaling solution on the rankings as described in Appendix 8.2 (showing that there is no obvious asymptote in the stress of a six dimensional solution over a seven dimensional solution).

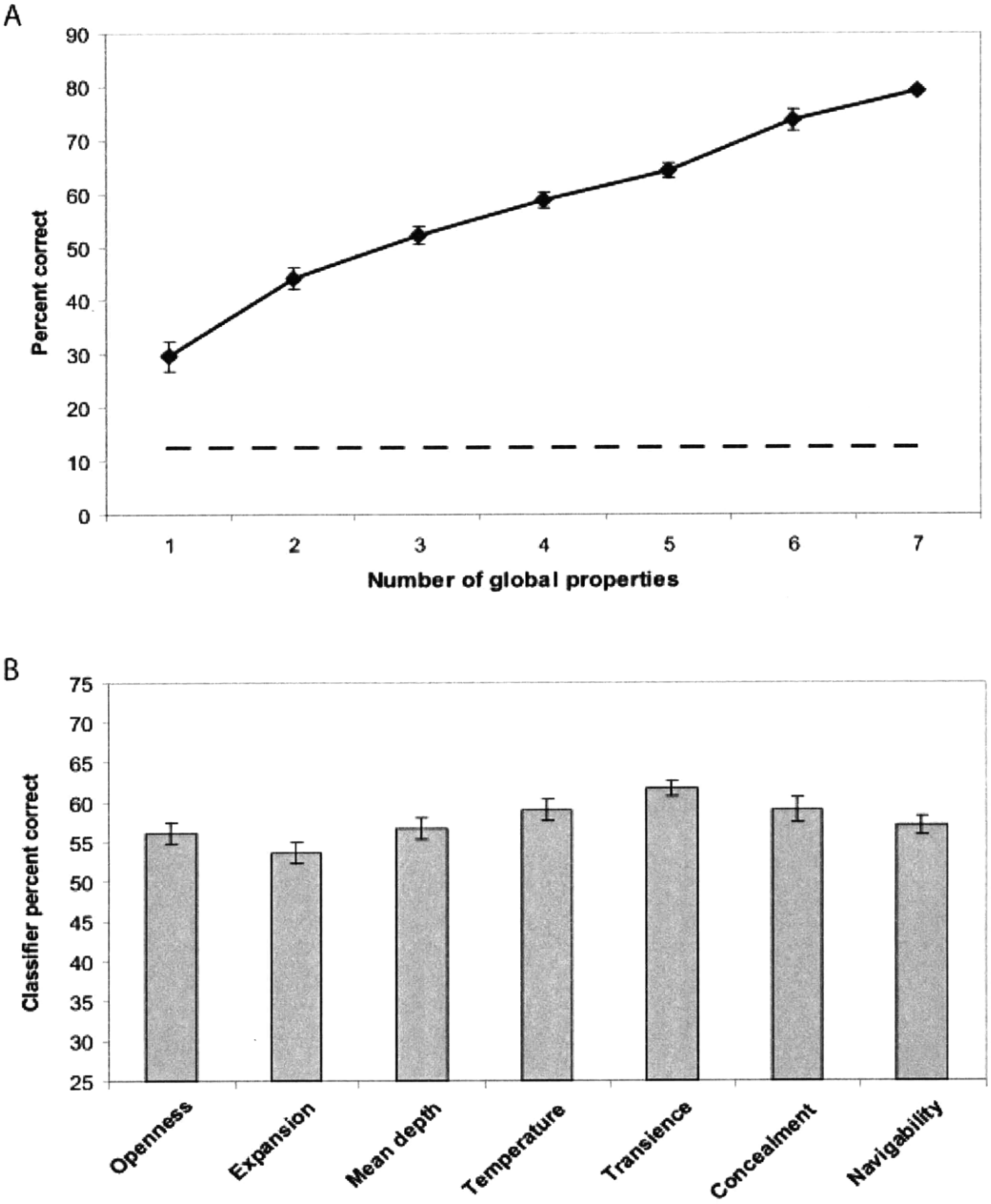


Figure 7: (A) Classifier's performance in Experiment 3 when trained with incomplete data, using from 1-7 global properties. The classifier can perform above chance with only one global property (30%), and performance linearly increases with additional properties. Chance level is indicated with the dotted line. (B) Mean classifier performance when trained with incomplete data that contained a particular global property. Classifier performed similarly when any particular global property was present.

5.26 – Global property classifier generalizes to less prototypical images

Up until this point, all scene images we have used have been ranked as being very prototypical for a basic-level scene category. However, scenes, unlike objects can often be members of more than one basic-level category (Tversky & Hemenway, 1983). A candidate scene representation is not complete without being able to generalize to and deal with images that span category boundaries. Many of the images in the natural world contain elements from multiple categories (poly-categorical). Take, for example the bottom image in Figure 8. This scene contains elements that could reasonably be assigned to forest, mountain, river or lake scenes. What assignment will the global property classifier give to such a scene?

Recall that the 200 typical scene images used so far were chosen from a larger pool of 500 images that had been ranked by human observers by how prototypical they were for these eight scene categories (Appendix 8.3). Recall also that the global property classifier is a maximum likelihood estimator, who computes the probability of an image being in each of the eight basic-level categories. Therefore, we can directly compare the order of category membership given by the human observers to the order of category probability given by the classifier (see examples in Figure 8).

First, for the 300 poly-categorical images, we compared the top-ranked choice from a category-ranking experiment (see Appendix 8.3) to the most likely category given by the classifier when trained on the 200 prototypical images. We found that the classifier's top category choice matched human observers' top category choice in 56% of images. It is of note that we would not expect the classifier performance on poly-categorical images to exceed its percent correct on prototype images (77%, section 5.21). It is also unreasonable to expect the model to agree better with human observers than these observers agree with each other about an image's category (Spearman's correlation 0.73, see Appendix 8.3).

A further complexity is that an image might be ranked equally prototypical for multiple categories, have possibility to be ranked in most categories, or have low overall prototypicality for all of the categories used in this experiment. In order to account for these, we then only analyzed images that received a score of at least 3 out of 5 for a category on the prototypicality scale (see Appendix 8.3 for method details), and those without a close second-place category rank. For these images, the model's top category choice matched the human observers' top category choice in 62% of the images. It is also notable that the top two category choices for the model match the top choice for the human observers in 92% of the images.

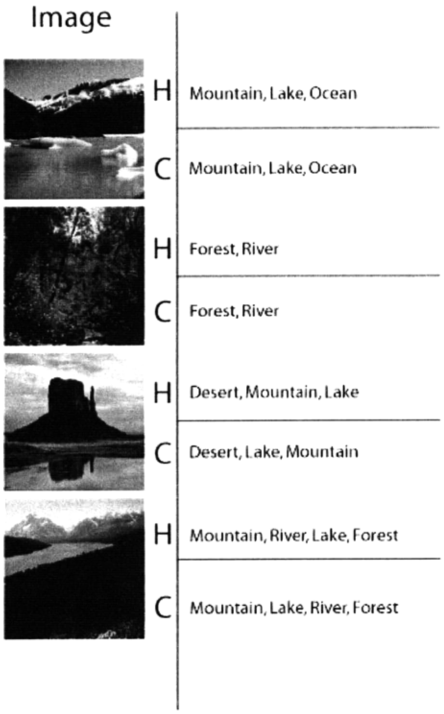


Figure 8: Examples of non-prototypical images. Human observers ranked the images according to their prototypicality along one or more categories (Appendix 8.3). For all

examples (H) indicates the order of prototypicality given by the human observers and (C) is the order of classification given by the global property classifier. Although the classifier rates the probability of the image being in each category, we show only the top choices for the same number of categories ranked by the human observers. In other words, if the human observers gave prototypicality rankings for two categories, we show the top two choices of the classifier.

5.3 – Discussion

Given that Experiment 2 showed that human observers were sensitive to global property information while rapidly categorizing natural scenes, in Experiment 3 we investigated the extent to which a scene-centered global description is sufficient to predict human rapid scene categorization performances. To do this, we employed a simple classifier whose only image information was the global property ranking data from Experiment 1. In terms of overall accuracy, the classifier is comparable to human performance (section 5.21), and has a similar performance by semantic category (section 5.21), indicating that the same semantic categories that are easier for human observers are also easier for the classifier. We have also shown that the errors made by the classifier are similar to the false alarms made by human observers (5.22-5.23). Critically, the exact errors are often repeatable (in other words, if a human observer makes a false alarm to a particular mountain as a forest, the classifier will most often make the same mistake). We have shown that the classifier, when trained on incomplete global property data, replicates the false alarms made by human observers in Experiment 2 when certain global properties were rendered less diagnostic for the classification task (sections 5.24 and 5.25). Finally, we have shown that the global property representation can deal with non-prototypical images as well as it deals with prototypical images (5.26). Altogether, we have shown that in terms of accuracy and errors, a representation that only contains global property information has high predictive value for human performance at rapid basic-level scene categorization.

6 – Experiment 4: An alternative hypothesis – comparing a global property representation to a local region representation

The global property based classifier shows remarkable human-like performance, in terms of both quantity and fidelity, in a rapid scene categorization task. Could any reasonably informative representation achieve such high fidelity? Basic-level scene categories are also defined by the objects and regions that they contain. Here, we test the utility of a local representation for predicting human rapid natural scene categorization by creating an alternative representation of our database that explicitly represents all of the local regions and objects in each scene. In order to fairly test the local representation, we employed two different models using these data, based on implementations of proposals in the literature: the local semantic concept model (Vogel & Schiele, 2007) and the prominent object model (Biederman, 1981; Friedman, 1979).

The *local semantic concept* model presents the case where an exhaustive list of scene regions and objects is created, and that scene recognition takes place from this list. Vogel and Schiele (2007) showed that very good machine scene classification could be done by representing a natural landscape image as a collection of local region names drawn from a small vocabulary of semantic concepts: an image could be represented as 9% sky, 25% rock, and 66% water, for example. Here we implement a similar idea, using the names of all regions and objects using a set of basic-level and superordinant region concepts along with their percent image area in a scene (see Method section 6.1 and Appendix 8.8 for details).

The *prominent object* model represents the case where scene recognition proceeds from a single, prominent object or region rather than an exhaustive list. This has been a popular potential mechanism for scene understanding proposed in the literature (Biederman, 1981; Friedman, 1979). Our implementation calculates the predictability of a scene category given the identity of the largest annotated object in the image. For example, we would predict that an image whose largest object is "trees" to be a forest, or an image whose largest region is "grass" is likely a field. Of course, objects can be prominent without necessarily being the largest objects, and a related literature is devoted to determining the image features that make an object prominent, or salient (for a review, see Itti & Koch, 2001). As the nature of these features is still relatively open, here we are limiting our definition of "prominent" to only include size.

It is important to note that both local region models present two conceptually different views about how scene recognition might proceed from local region and object information. The *local semantic concept* model categorizes a scene based on the co-occurrence of regions from an exhaustive list of scene regions, assuming that in a glance all objects can be segmented, perceived and abstracted into concepts. This model represents the best-case scenario for the local approach, in which the identities of *all* of the objects and regions in the scene are known, as well as their relative sizes.

By contrast, the *prominent object* model assumes that not all regions have equal diagnostic information for the scene category, and that in particular, if an object is prominent in the scene, it will contain more information about the scene's category.

Scene categorization is therefore an inference based on the recognition of this prominent (and informative) object. However, it is important to note that as size information is also included in the local semantic concept model, all of the information in the prominent object model is contained in the local semantic concept model. Therefore, the essential difference in the two models is in the relative importance of one object verses the importance of all objects.

6.1 – Method

Two independent observers (one author, and one naïve observer) hand-segmented and labeled all regions and objects in the 200 image database. The labeling was done using the online annotation tool LabelMe (Russell, Torralba, Murphy & Freeman, 2008). Example annotations are found in Figure 9. There were a total of 199 uniquely labeled regions in the database. All of the labels were pared down to 16 basic and superordinant level region names by removing typos, misspellings, synonyms and subordinant-level concept names (for example "red sand" instead of "sand"). We used the following region concepts for the local semantic concept model: *sky, water, foliage, mountains, snow, rock, sand, animals, hills, fog, clouds, grass, dirt, manmade objects, canyon* and *road.* This list includes the nine semantic concepts used by Vogel & Schiele (2007) as well as others that were needed to fully explain our natural image database. In Appendix 8.8, we report that the performance of this 16 concept model is not different from a model using the raw data (199 concepts), or a model using 50 basic-level region concepts.

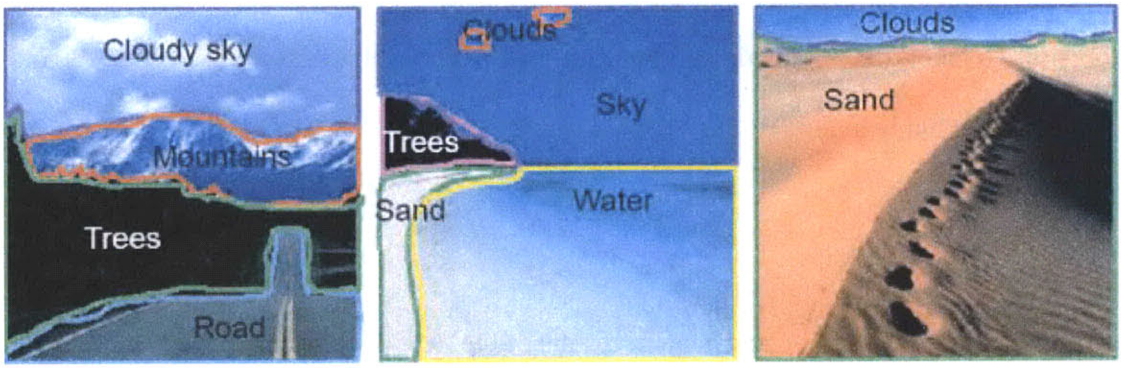


Figure 9: Examples of segmentations and annotations made using the LabelMe annotation tool, and used as the basis for the local scene representation in Experiment 4.

Each image's list of regions (along with their image area) was used to train and test a naïve Bayes classifier using the same leave-one-out procedure as described in Experiment 3. As with the global property classifier of Experiment 3, results are compared to the human psychophysical performance of Experiment 2.

For the *prominent object* model, the naïve Bayes classifier was not needed because the relevant information could be calculated directly from the statistics of the LabelMe annotations. For each image, we calculated the probability of the image being from each basic-level category based on the identity of the scene's largest object. For this analysis, we used the 50 local concept list (see Appendix 8.8) as it had the best balance between distinctiveness and representation of the object concepts.

For each image, we computed a 50 region by 8 category matrix of object predictability from the 199 remaining scenes where each entry (i,j) was the probability of the region (i) being in the basic-level category (j). Taking the row representing the largest region in the test image, we selected the category for maximum probability for that region. For example, if a scene's largest region was *sky*, the probabilities of the scene being from each of the eight categories are as follows: 0.20 desert; 0.14 field; 0.04 forest; 0.16 lake; 0.17 mountain; 0.15 ocean; 0.05 river; 0.09 waterfall. Therefore, the scene is most likely a *desert*.

6.2 – Results

A summary of the classification results of the two local region models, along with a comparison to the global property model of Experiment 3, can be found in Table 5.

	Percent correct	**By-category correlation**	**Item analysis correlation**	**Between-category confusion correlation**
Prominent object model	52%	0.55	0.69*	0.06
Local semantic concept model	60%	0.64	0.69*	0.23
Global property model	77%	0.88*	0.76*	0.77*

Table 5: A summary of performance of local region-based models tested in Experiment 4 with the global property model of Experiment 3. The *local semantic concept* model refers to a model in which a scene is represented as a co-occurrence vector of all labeled regions and objects along with their relative sizes. The *prominent object* model refers to the predictability of the scene category conditioned on the presence of its largest object. The by-category correlation (cf. section 6.21 for the local models and 5.21 for global model) shows the extent to which the models are similar to the pattern of human correct performance rate by category for the eight basic-level categories. The item analysis (section 6.22 and 5.22 for local and global models respectively, bins of 25) shows the extent to which the models tend to misclassify the same images as humans do. The between-category confusion correlation (section 6.23 and 5.23 for local and global models respectively) shows the extent to which the patterns of confusability between pairs of basic-level categories for the models were similar to those of human observers. (*) indicates significant correlations (p<0.05).

6.21 – Local models' performance: Percent correct and correlation to human basic-level category performance

The *local semantic concept* model averaged an overall 60% correct categorizations (vs. 77% for the global property classifier, section 5.21), which was significantly lower than the percent correct of human meta-observers (77%, t(7)=-2.88, p<0.05, also recall 86% for complete observers). To ensure that the local semantic concepts were not too general, we compared this performance to the performance on a larger list of 50 basic-level region concepts, finding no significant performance difference to the semantic concept model (t(398)<1, see Appendix 8.8 for details, including the percent of correct classifications per semantic category). The *prominent object* model performed well overall. The overall percent correct for this model was 52% (chance being 12.5%), but still under the rate of human observers (t(7)=-9.4, p<0.0001).

To evaluate how the local models compared to human performance by category, we correlated meta-observer correct performance and object models' correct performance for the 8 basic-level categories (as in Section 5.21 and Figure 5 for the global property model): None were significant (r=0.64, p=0.09, for the *local semantic* model, and r=0.55, p=.16, for *prominent object* model).

These results suggest that the scene categories that are easy or hard for human observers to classify at a short presentation time are not necessarily the same for the objects models. In fact, the categories *field*, *forest* and *mountain* are classified by all three models at human performance levels, whereas the object models' classifications drop for *desert*, *lake*, *ocean* and *river*. Indeed, *field*, *forest* and *mountain* are environments that are mostly composed of one or two prominent regions or objects (e.g. grass for field, trees for forest, and mountain for mountain), whereas other scene categories share more objects between them, putting local models at a disadvantage.

6.22 – Error analysis: Easy and Difficult Images

As we did in Experiment 3 (section 5.22), we performed an item analysis to determine if the local region models would have trouble classifying the same images that human observers do. This analysis quantifies whether an error is made on an image, but not the type of error made.

Both the local semantic concept model and the prominent object model reflected the level of difficulty of the images for humans as well as the global property model did (for bins of 25, r=0.69 for both object models, both correlations significant p<0.001, see Table 5. Bins of 10 yielded higher coefficients, r=0.89 for local semantic concept model and r=0.85 for the prominent object model). These correlations indicate that both global and local representations have a tendency to perform well or poorly on the same images. However, this analysis does not give information about the type of errors made. In other words, the local models and human observers tend to misclassify the same images, but do they misclassify these images as being the same category? We explore this issue below.

6.23 –Qualitative error analysis: Distribution of error types

In order to evaluate further the types of errors made by the local models, we analyzed the extent to which the distribution of errors made by the object models was similar to the distributions of false alarms made by human observers. For instance, in the rapid scene categorization task (Experiment 2), humans often confused *river* and *waterfall*, as well as *desert* with *field* (Table A5). However, they almost never mistake a *forest* for an *ocean*. Are the pairs of categories often confused by human observers also often confused by the local region models? As in section 5.23, we compared the pairwise basic-level category confusions made by the local region models to the distribution of false alarms made by the human observers for each pair of categories. For both local models, there was no significant relation between their patterns of category confusability and those of the human observers: r=0.23 (p=.25) for the *local semantic concept model*, and r=0.06, (p=0.75) for the *prominent object* model (the global property model gave r=0.77 for comparison). This indicates that there is limited similarity between the local models and human observers in terms of the pairs of categories confused, and suggests that these local models do not capture the richness of the representation built by human observers in a 30 msec presentation time.

6.3 – Discussion

The high performance of the global property model begs the question of whether any reasonably rich and informative representation could predict human rapid scene categorization performance.

Here we have explored two distinct alternative hypotheses to the global property scene representation. In particular, our results suggest that a local, region-based approach, based on suggestions from the literature does not have the same capacity to explain human rapid scene categorization as the global property model does. It is of note that the *local semantic concept* model represents one of the best-case scenario for the local approach, in which the identities of *all* of the objects and regions in the scene are known, as well as their relative sizes.

While the local semantic concept model shows relatively good percent correct performance at basic-level scene categorization (60%, chance being 12.5%), it does not have the fidelity to predict the types of false alarms made by human observers in a rapid scene categorization task (c.f. Table 5). For instance, Figure 10 shows example false alarms made by the global property classifier of Experiment 3 with the local semantic concept model of Experiment 4. Strikingly, the top desert and river are classified by the global property classifier as being field and forest respectively. This mirrors the pattern of false alarms made to the same images by human observers in Experiment 2. However, the lake and river shown at the bottom of Figure 10 were classified as ocean and field respectively by the local semantic concept model; errors that were not made often by the human observers in Experiment 2. At first glance, it seems strange that such a prototypical river (bottom right of Figure 10) would be classified as a field at all. However, as fields in our database have large amounts of sky, trees and rock (similar to rivers), this image was classified as a field by the local semantic concept model.

The prominent object model, while having the lowest overall correct categorization performance of the models, still performed substantially above chance. This is because some categories, such as *field* and *forest* were very well categorized by this model. This makes intuitive sense, as typical prominent objects for these categories were *grass* and *trees* respectively, which were very diagnostic for these categories. However, these categories which were easy for the model to classify had limited similarity to the categories that were easy for the human observers to classify, which is why the by-category correlation was modest. While the prominent object model had a tendency to correctly categorize the same images as human observers, it could not predict the types of errors that the human observers would make. For example, if water was the largest object in a scene, the prominent object model could not distinguish whether the scene was a lake, ocean, river or waterfall because water is equally diagnostic for these categories.

Likewise, the local semantic concept model was able to correctly classify the majority of the images in the database. This is because there is a considerable amount of redundancy in image categories that allowed the model to learn that a scene with *cliffs*, *water* and *sky* is likely to be a waterfall while a scene with *sand*, *rock* and *sky* is likely to be a desert. However, the pattern of correct category classification of this model showed only modest similarity to that of the observers. For example, *field* was very well classified by the model while it was on average, one of the more difficult categories for the human observers in the rapid categorization task. This is likely because the model was relying heavily on the presence of objects such as *grass* or *flowers* that are unique to this category. Like the prominent object model, the local semantic concept model tended to correctly classify the same images as human observers, but could not predict the types of false alarms made by humans. In particular, categories such as *lake* and *river* have very similar sets of objects (typical objects include *sky*, *water*, *trees* and *grass*), so it was difficult for the local semantic concept model to distinguish between these categories, even though human observers did not have such a difficulty.

In contrast, the global property model of Experiment 3 had higher correct classification performance than the local models, and was very similar to human observers' performance. Also in contrast to the local models, its pattern of performance by category significantly correlated with that of the human observers'. Like both of the local models, it also tended to correctly classify the same images that human observers did. However, unlike the local models, it has the power to predict the types of false alarms made by the human observers. To go back to the *lake* and *river* example, the local models made errors in these categories because the objects in them are very similar.

However, the global property model can distinguish between them because they have different layout and surface properties: lakes are more open, and less transient, for example (see Figure 3). To the human observers, few errors are made between these categories, perhaps because the observers are using the structural differences between these categories to distinguish them.

Clearly, more sophisticated object models that incorporate structure and layout information should be able to capture more of the essence of a natural scene (Grossberg & Huang, in press; Murphy, Torralba, Freeman, 2003). Our point here is that object models testing simple instantiations of valid propositions from the visual cognition literature do not have the same explanatory power as our global property model for predicting human rapid scene categorization performance.

Importantly, we do not mean to imply that local objects are regions are not represented in early processing of the visual scene. Instead we have shown that the remarkable fidelity of a global property representation for predicting human rapid scene categorization performance cannot be achieved with any reasonably informative description of the visual scene.

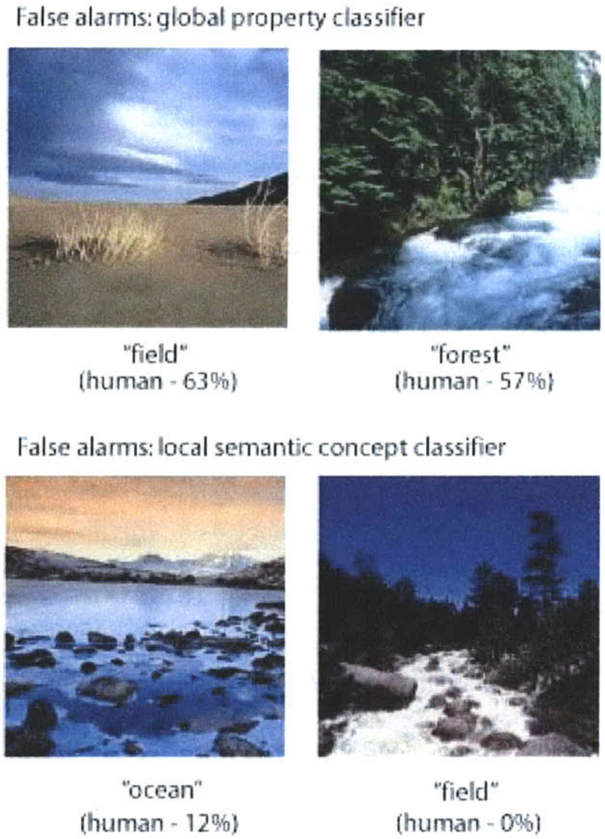

False alarms: global property classifier
"field"
(human - 63%)
"forest"
(human - 57%)
False alarms: local semantic concept classifier
"ocean"
(human - 12%)
"field"
(human - 0%)

Figure 10: Examples of false alarms made by the global property classifier of Experiment 3 and the local semantic concept classifier of Experiment 4. Underneath, we report the percent of human false alarms made on that image. The global property classifier captures the majority of false alarms made by human observers while the local semantic concept classifier captures less (see Table 5).

While local region and object information most certainly make up an important part of a scene's identity, our results suggest that the representation formed by human observers after a very brief glance at a scene is not dominated by local object information (see also Fei Fei et al, 2007). Our results suggest the possibility that our qualia of object perception in a brief glance might be based upon inference of these objects given global scene structure and schema activation.

7 – General discussion

In this work, we have shown that a global scene-centered approach to natural scene understanding closely predicts human performance and errors in a rapid basic-level scene categorization task. This approach uses a small vocabulary of global and ecologically relevant scene primitives that describe the structural, constancy and functional aspects of scene surfaces without representing objects and parts. Beyond the principle of recognizing the "forest before the trees" (Navon, 1977), here we propose an operational definition of the notion of "globality" for natural scene recognition, and provide a novel account of how human observers could identify a place as a "forest", without first having to recognize the "trees".

Several independent analyses, on human performance alone (Experiments 1 and 2), and on human performance compared to a classifier (Experiments 3 and 4), were undertaken to finely probe the relation between a global scene representation and human rapid natural scene categorization performance. Although strict causation cannot be inferred from these correlational results alone, all results taken together are suggestive of the view that a scene-centered approach can be used by human observers for basic-level scene categorization. Strengthening this view is the fact that performance of a classifier representing the local objects and regions of the images (Experiment 4) does not have the same explanatory power as the global property representation (Experiment 3) for predicting human performance and false alarms (Experiment 2).

We have shown that human performance at a rapid scene categorization task can be dramatically influenced by varying the distractor set to contain more global property similarities to a target category (c.f. Figure 4, section 4.22). Moreover, the item analysis which calculates the probability of a false alarm occurring to single distractor images, was very well predicted from each distractor's distance from the target-category mean for a global property, suggesting that rapid image categorization performance follows the statistical regularities of global properties' distributions in basic-level categories. Last, the relative confusability of basic-level categories (section 4.23, Tables A5 and A6) to one another is also well-explained by the basic-level categories' similarity in global-property space.

To determine how computationally sufficient the global properties are for explaining the human rapid scene categorization data in Experiment 2, we compared a simple classifier to human performance on several metrics (Experiment 3). First, the overall categorization performance of the classifier was similar to humans', and the relative performance of the classifier by category was also well correlated with human observers.

However, similar levels of performance are not enough: if the global property representation is a plausible human scene representation, then the classifier should also predict the false alarms made by human observers. We have shown that image difficulty for the classifier is very similar to image difficulty for human observers, and that the same qualitative errors are made by both (e.g. false alarming to a particular *river* image as a *waterfall*) the majority of the time (sections 5.23). Furthermore, we have shown that when a global property is not available for use in categorization, either because it is not explicitly represented (classifier), or because the distractors make it non-diagnostic of the target category (humans), performance suffers similarly (sections 5.24-5.25). Furthermore, we have shown in section 5.26 that the high fidelity of categorization performance in the global property model can generalize beyond prototypical images. In particular, the level of agreement between the classifier and human observers is not different from the agreement between the human observers. Lastly, the striking predictability of the global property model for human scene categorization performance is not found in two local object models that we tested (Experiment 4).

It has been known that visual perception tends to proceed in a global-to-local manner (Navon, 1977), but for stimuli as complex as a natural scene, it is not obvious what the global level might be. Computational models have shown that basic-level scene categories can emerge from a combination of global layout properties (Oliva & Torralba, 2001, 2002, 2006), or from a collection of regions (Fei Fei & Perona, 2005; Grossberg & Huang, in press; Vogel, Schwaninger, Wallraven & Bulthoff, 2006; Vogel & Schiele, 2007) but no psychological foundation has yet been established between global scene properties and basic-level scene categorization performance. This work has tried to make this link. By grounding our search in the principles of environmental affordance (Gibson, 1979; Rosch, 1978), we found a collection of global properties that are sufficient to capture the essence of many natural scene categories.

Our result is also in the spirit of seminal scene understanding studies from the 1970s and 1980s. Biederman and collaborators have shown that coherent scene context aided the search for an object within the scene, even when the identity and location of the object were known in advance (Biederman, 1972). Furthermore, lack of coherent spatial context seemed particularly disruptive on negative trials where the object was not in the scene, but had a high probability of being in the scene (Biederman, Glass & Stacy, 1973). Together, this suggests that scene identity information may be accessed before object identity information is complete. Biederman (1981) outlined three paths by which such scene information could be computed: (1) a path through the recognition of a prominent object; (2) a global path through scene-emergent features that were not defined at this time; (3) the spatial integration of a few context related objects.

Our results offer positive evidence for path 2 (the global path suggested by Navon, 1977, but never operationalized) and non-conclusive evidence for path 1 (the prominent object). Path 3 supposes that the co-occurrence of a few objects in a stereotypical spatial arrangement would be predictive of the scene category. The semi-localized local model of Vogel & Schiele (2007) along with the studies of relation processing by Hummel and colleagues (e.g. Saiki & Hummel, 1998) has started to find evidence for this path. However, there is also reason to believe that path 3 may not be the only approach for capturing the type of representation built over a brief glance at a novel scene. This view requires that several objects be segmented, recognized and relationally organized for scene categorization to occur. However, it is still not clear that humans can segment, identify and remember several objects in a scene at a glance. Potter et al. (2004) demonstrated that, in a memory test following an RSVP sequence of images, a large number of false alarms were made to images that were conceptually similar to an image presented in the sequence, but did not necessarily have the same objects and regions, suggesting that what is encoded and stored from a brief glance at a scene is a more general description of the image than an exhaustive list of its objects. This view is corroborated with the facts that human observers also make systematic errors in remembering the location of objects from a briefly glimpsed display (Evans & Treisman, 2005), and are relatively insensitive to changes in single objects in a scene (change blindness, Rensink et al, 1997; Simons, 2000).

A consequence of our global precedence finding could be that the perceptual entry-level for visual scenes is not the basic-level category, but rather an image's global property descriptions, at a superordinate level (Joubert et al., 2007; Oliva & Torralba, 2001, 2002). This idea is not necessarily contradictory of the behavioral findings of Rosch and colleagues. We argue that the basic-level category is the entry level for *communication* about objects and places because it represents a compromise between within-category similarity and between-category distinctiveness. However, under the constraints of a rapid categorization task, perhaps the initial scene representation would benefit from processing *distinctiveness* first, making a superordinate description an ideal level, particularly if the visual features used to get this superordinate description do not require a segmentation stage, known to be computationally more expensive than an holistic analysis (Oliva & Torralba, 2001).

Finding the image-level features that mediate such rapid visual categorizations is a fascinating, yet rather open question that is beyond the scope of the current work. Indeed, previous work has shown that certain spatial layout properties, such as *openness* and *mean depth* can be well-described from a set of low-level image features corresponding to spatially localized second-order image statistics (Oliva & Torralba, 2001, 2002; Torralba & Oliva, 2002, 2003). Some properties, such as *temperature*, might even be represented by simpler images features, such as the color distribution. However, functional properties such as *navigability* and *concealment* may be more complex to represent, as their spatial structures might not co-vary in a simple way with first or second order image statistics. For instance, if a scene is very *open*, it is open because it has a very salient horizon line somewhere near the vertical center, and all scenes that are consistently ranked as *highly open* share this feature. A *navigable* scene however, might be navigable because the scene is open and free of clutter, or it could be navigable because it has a very obvious path through an otherwise dense environment. Therefore, image features of a higher complexity might be needed to fully represent these global properties, a question that future research will investigate.

A global scene-centered representation is a plausible coding of visual scenes in the brain and a complementary approach to object-based scene analysis. This present work suggests that rapid scene recognition can be performed by global scene-centered mechanisms and need not be built on top of object recognition. Indeed, work in functional imaging has shown a dissociation between brain areas that represent scenes (the parahippocampal place area, or PPA, Epstein and Kanwisher, 1998) and those that represent individual objects (Bar, 2004; Grill-Spector, Kourtzi & Kanwisher, 2001). Furthermore, the PPA seems to be sensitive to holistic properties of the scene layout, but not to its complexity in terms of quantity of objects (Epstein and Kanwisher, 1998). The neural independence between scenes and object recognition mechanisms was recently strengthened by Goh, Siong, Park, Gutchess, Hebrank & Chee (2004). They observed activation of different parahippocampal regions when pictures of scenes were processed alone compared to pictures containing a prominent object, consistent within that scene. Steeves, Humphreys, Culham, Menon, Milner & Goodale, (2004) have shown that an individual with profound visual form agnosia could still identify pictures of real world places from color and texture information only. These findings are consistent with the hypothesis that whole scene recognition may be dissociated from object identification.

What is the mechanism by which a scene-centered pathway could arise in the brain? Although we are far from a definitive answer, an examination of the time course of visual processing yields critical insights. Thorpe and colleagues (1996) have made a case that the speed of high-level visual processing necessitates a single feed-forward wave of spikes through the ventral visual system. Furthermore, biologically inspired models of this architecture yield high performances in detection tasks (Delorme & Thorpe, 2003; Serre, Oliva & Poggio, 2007). However, very rapid feedback might also mediate this performance. Physiological evidence shows that there is considerable overlap in time between spikes arriving in progressive areas of the ventral visual stream (Schmolesky, Wang, Hanes, Thompson, Leutgeb, Schall & Leventhal, 1998), suggesting that feedback from higher visual areas can feed back to early visual areas to build a simple yet global initial scene representation. Furthermore, a combined EEG/MEG and fMRI study has shown a V1 feedback signal as early as 140msec after stimulus presentation (Noesselt, Hillyard, Woldorff, Schoenfeld, Hagner, Jancke, Tempelmann, Hinrichs, & Heinze, 2002) furthering the idea that scene recognition may be mediated through rapid feedback. Strikingly, there is evidence of the global pattern from a contextual cueing display being processed 100 msec after stimulus presentation (Chaumon, Drouet & Tallon-Baudry, 2008). These results confer with behavioral evidence which suggest that global properties such as *concealment* or *naturalness* are available for report with less exposure time than basic-level categories (Greene & Oliva, in preparation; Joubert et al, 2005, 2007; Kaplan, 1992). Although this does not necessarily imply that they are processed first by the brain, it is consistent with the view that global properties are reasonable scene primitives for basic-level categorization.

Emphasizing the importance of a scene-centered view does not imply that objects are not an important part of rapid scene recognition. Surely, as objects can make up the identity of the scene and are the entities acted on by agents in a scene, they are of critical importance for scene understanding with longer image exposures. However, it appears that objects might not necessarily be the atoms of high level recognition especially under degraded conditions of blur or at the very beginning of visual analysis (Oliva & Schyns, 2000; Schyns & Oliva, 1994). But given longer image exposures, objects become increasingly important in our representations of scenes during the course of the first fixation (Fei Fei et al, 2007; Gordon, 2004) and a framework that would combine objects and their spatial relationships with global properties would capture more of the richness of scene identity.

In this work, we have demonstrated that global property information is more diagnostic of natural scene categories than local region and object information. A natural question is then what roles both types of information play in other types of environments, such as indoor scenes? Intuitively, the prominent object model from Experiment 4 seems like it would do a good job at categorizing some indoor categories such as bedrooms or living rooms because the largest object (bed or sofa) is not typically found in other scene categories. However, it does not seem that all indoor categories are so strongly object-driven. A corridor, for example, is unique among indoor scene categories as having a great deal of perspective. A conference room and a dining room might also be confused by a prominent object model as they both have prominent tables surrounded by chairs. Part of our ongoing effort is characterizing the relative use of global and local diagnostic information for scene categorization for a greater variety of scene categories.

An extension of the present work that could indirectly probe the neural representation of visual scenes is to measure if global properties are adaptable (Greene & Oliva, 2008). A ubiquitous property of neural systems is that repeated presentation of a represented property leads to a temporary decrease in sensitivity to that property, a phenomenon known as adaptation. This phenomenon is seen at all levels of visual processing for entities that seem to have dedicated processing, from basic properties such as color, motion, orientation and spatial frequency (for a review, see Wade & Verstraten, 2005) to complex features such as facial emotion and identity (Webster, 2004; Leopold, O'Toole, Vetter & Blanz, 2001). Furthermore, adapting to low-level image features can modulate higher level perceptual judgments for surface glossiness (Motoyoshi, Nishida, Sharan & Adelson, 2007) or the naturalness of real-world scenes (Kaping, Tzvetanov & Treue, 2007).

7.1: Concluding remarks

The present work was designed to operationalize the notion of globality in the domain of natural real-world images. We have shown that global properties capture much of the variance in how real world scenes vary in structure, constancy and function, and are involved in the representation of natural scenes that allows rapid categorization.

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

8 – Appendix

8.1 – Pilot experiment for determining global properties

In order to ensure that image properties and affordances stated in the literature are relevant to our natural scene image database and participant population, we ran the following pilot experiment with 5 naïve observers. Participants viewed each of the 200 natural landscape images, one at a time for one second each. Observers were given the following instruction: "We are studying how people perceive space in photographs. Describe the kinds of actions that you could do if you were in that scene at that moment, from that viewpoint. You might also mention what you might not be able to do due to environmental conditions". Observers typed their answer in a free-response prompt, and were given unlimited time.

Observers' responses were tabulated by one author as to the broad environmental concepts they contained. Table A1 summarizes these concepts (see caption for details). Recognizing the possibility for experimental bias in this method, care was taken to be as conservative with tabulations as possible. The descriptors given are similar to those found in other studies of environmental interaction (Appelton, 1975; Kaplan, 1992), and of environmental spatial layout (Oliva & Torralba, 2001). All of the global properties used in the subsequent experiments (*openness, navigability, mean depth, concealment, perspective, transience,* and *movement*) were conceptually mentioned or described by all participants.

Concept	Mean frequency mentions per image
Navigation	**1.39 (5)**
Exploration	**0.26 (5)**
Temperature	**0.17 (5)**
Movement	**0.15 (5)**
Space	**0.14 (5)**
Camouflage	**0.12 (5)**
Harvest	**0.11 (5)**
Rest	**0.06 (4)**
Water	**0.06 (3)**
Animal	**0.03 (2)**
Ruggedness	**0.02 (2)**

Table A1: Mean mentions of scene properties per image in the scene description study (see Appendix 8.1). The number in parentheses indicates the number of observers who have mentioned the concept (out of 5 total observers). *Navigation* refers to self-propelled land or water movement through the scene (e.g., walking, running, swimming, driving). *Exploration* refers to examination or interaction with a particular object (e.g. look at, play with). Although this was mentioned by all participants, it was not included as a global property because it refers to interactions with single objects, and not the entire scene. *Temperature* contains references to the physical temperature of the environment (e.g. hot, cold, warm). *Movement* refers to statements of the scene in change or anticipation of it changing ("wait for car", "water is too fast to swim"). This is a similar concept to *transience* in Experiments 1, 2 and 3. *Space* includes mentions of the size or physical geometry of the scene (*openness*, *perspective*, *mean depth*). *Camouflage* contains references to either the human being able to hide in the scene or that something/someone could be hidden in the scene ("hide in trees", "watch for birds"). This is a similar concept to *concealment* from Experiments 1, 2 and 3. *Harvest* contains references to taking something from the environment (e.g. picking flowers, hunting and fishing). *Water* refers to the presence of, or search for water. *Rest* contains repose words such as "sit" or "lie down". *Animal* contains references to animals that are either present in the scene or could potentially come into the scene. *Ruggedness* contains references to aspects of the environment that make navigation treacherous.

8.2 – Global property space

In the ranking task of Experiment 1, there was considerable spread in the ranking values for each of the basic-level categories (waterfall, river, ocean, mountain, lake, forest, field and desert) along each global property (see Table 1). Figure A1 shows every image's rank for each global property, broken down by basic-level category (see the Method section of Experiment 1).

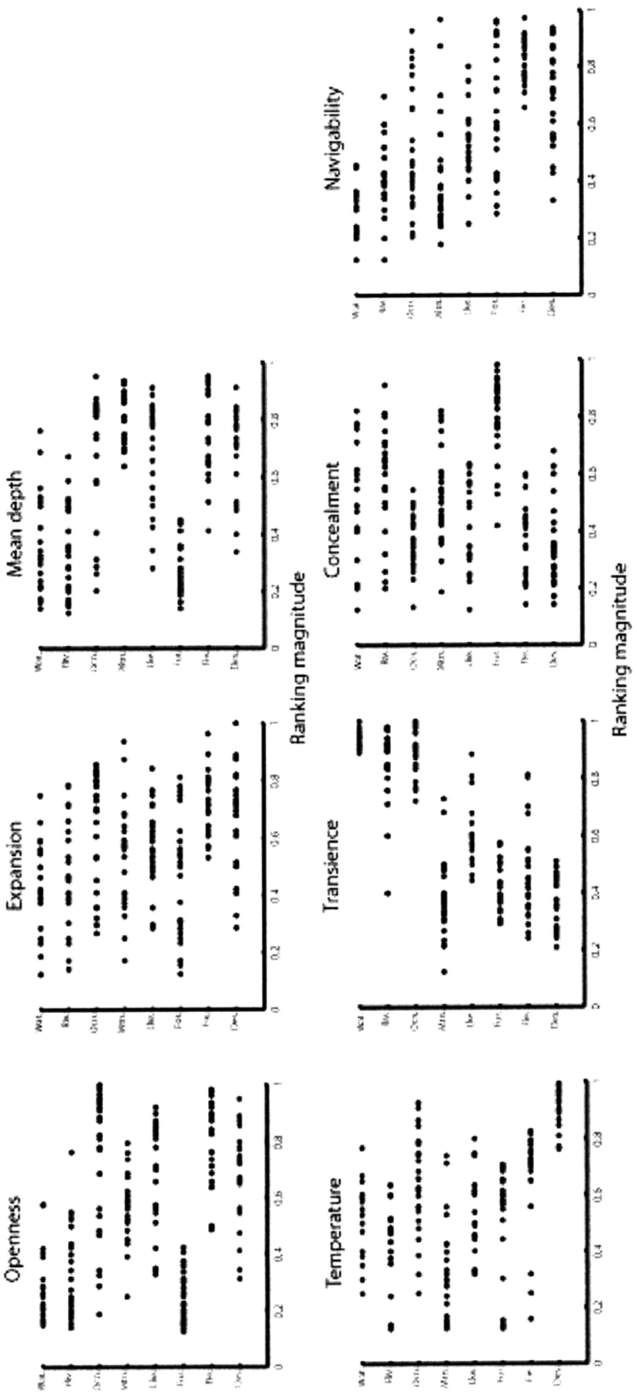


Figure A1: The figure shows the mean rank of each of the 200 scene image, in their respective semantic category, along each of the seven global properties. These are from the ranking data from Experiment 1. In all basic-level categories, there is a considerable spread of image rankings, indicating that the eight basic-level categories used in Experiment 1,2,3 and 4 do not cluster along single global properties. Abbreviations of the basic-level categories correspond to: Waterfall, River, Ocean, Mountain, Lake, Forest, Field and Desert.

Table A2 shows the correlations between the images' ranking along one global property to the images' ranking along each other global property, from Experiment 1. Correlations between image rankings were computed for each pair of global properties in the database.

	Openness	Expansion	Mean depth	Temperature	Transience	Concealment	Navigability
Openness	*						
Expansion	**0.75**	*					
Mean depth	**0.90**	**0.70**	*				
Temperature	**0.35**	**0.29**	0.19	*			
Transience	**-0.22**	**-0.22**	**-0.34**	-0.13	*		
Concealment	**-0.52**	**-0.24**	**-0.43**	-0.17	-0.06	*	
Navigability	**0.53**	**0.64**	**0.40**	**0.46**	**-0.44**	0.13	*

Table A2: Correlations between pairs of global properties (image by image) from the human ranking data of Experiment 1. Correlations that are statistically significant are shown in bold.

It is of note that these correlations are more a reflection of the landscape images in the natural image database we used, and less a statement about the similarity of the property concepts. For example, in this database *openness* and *mean depth* are highly correlated. However, previous work has shown that for a larger and more diverse database of real world scenes, this relation is much less strong (Oliva & Torralba, 2002).

While the global properties are not all statistically independent with each other (Table A2), each property gives unique information about the scene images. For example, while all *open* places also have large *mean depth*, not all large depth pictures are necessarily open (see Figure A2-a). Likewise, places that are easily *navigable* might or might not be have *perspective* (see Figure A2-b), and two very closed places such as forests can have different degree of expansion (see Figure A2-c). It's of note that *concealment* and *navigability* are not correlated with one another (*r*=0.13). This is because it is the size and distribution of the obstacles in a scene that matter for estimating these properties in a given space, and not merely the presence of obstacles. For example, a very dense forest of thin trees does not provide good cover for a human (low navigability and low concealment), and a forest with a clear path through it would rank highly for both navigability and concealment.

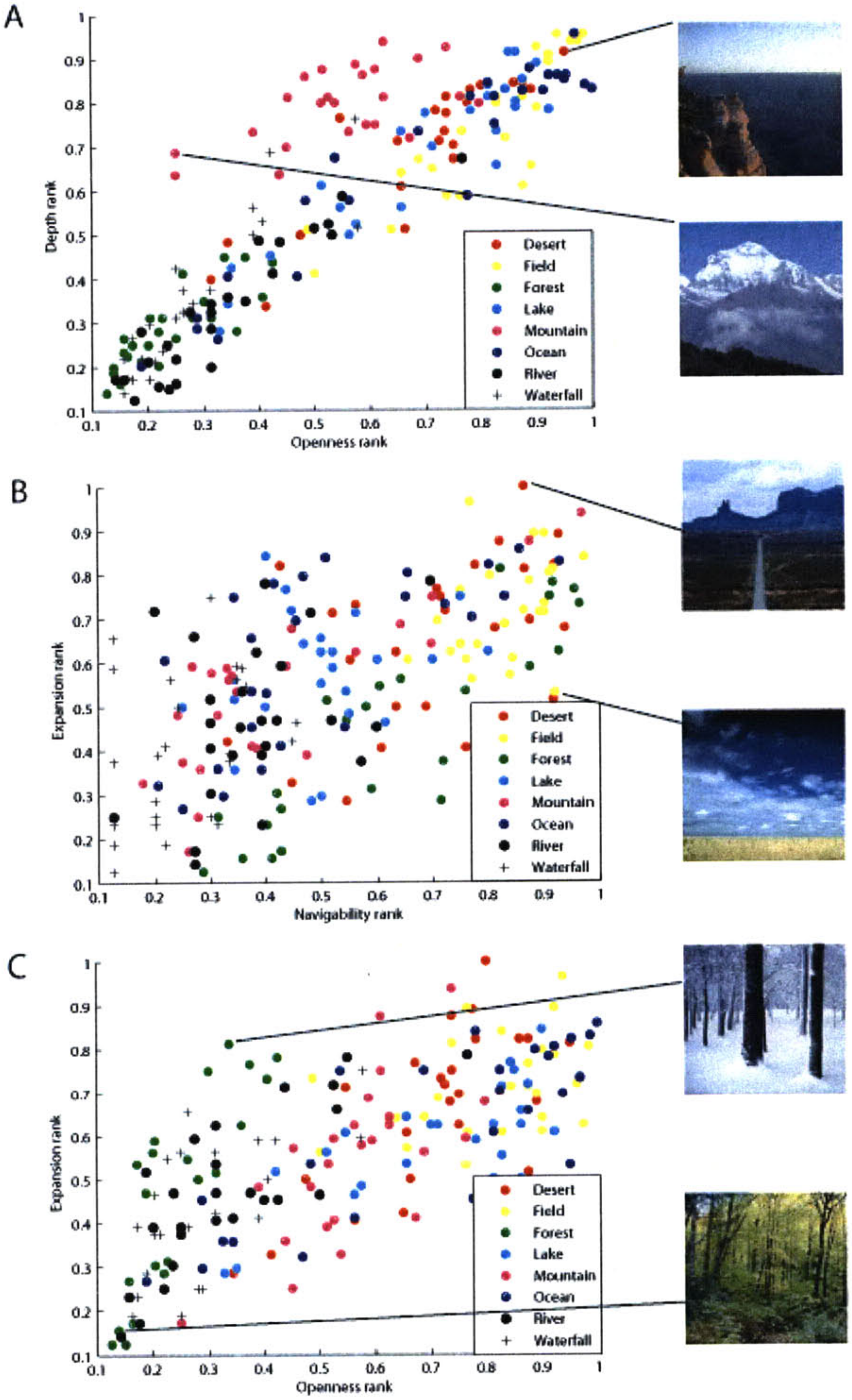


Figure A2: A) A scatterplot of the rankings of the 200 natural scenes along *mean depth* and *openness* (from Experiment 1) shows that although there is a strong correlation between these properties in this particular database, these properties represent distinct spatial concepts. For example, images with *large depth*, can either be very *open*, with an infinite horizon like the picture of the canyon, or moderately *closed* such as the mountainous landscape scene, where the horizon is bounded by a peak. B) A scatterplot showing all image ranks along the *navigability* and *expansion* dimensions. The two images shown are

perceived as having a *high degree of navigability*, however they have a different linear perspective; C) A scatterplot between *openness* and *expansion* dimensions, illustrated the fact that *open* environments may have different degree of perspective. Each dot in the scatterplot represents the mean rank of one image, averaged over at least 10 observers.

To further test the structure and dimensionality of the ranking data of Experiment 1, we employed classical multidimensional scaling (MDS) from the Euclidean distance matrix of images along the seven global properties. The first three dimensions of the solution are plotted in Figure A3-a. The eigenvalues of the y*y' transformation matrix are plotted in Figure A3-b. Unfortunately, there is no objective test of MDS dimensionality. A "scree" or elbow test is typically employed to test the underlying dimensionality of an MDS solution. The lack of an obvious elbow as shown in Figure A3-b suggests that all seven dimensions, although correlated, contribute to the scene category representation.

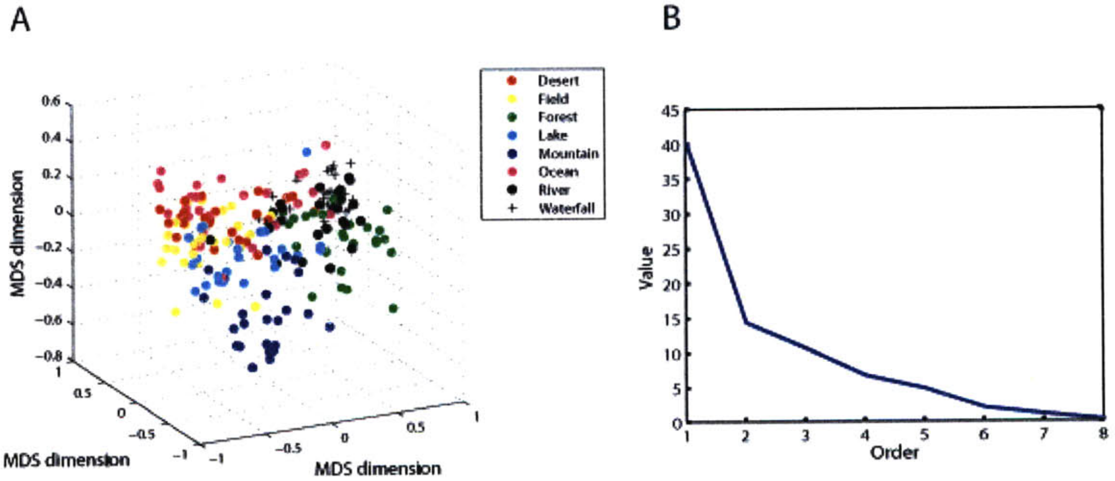


Figure A3: The classical multi-dimensional scaling (MDS) solution for the global property rankings from Experiment 1. a) A scatter plot of each of the 200 scenes in the database projected onto the first three MDS dimensions. Different semantic categories are shown in different colors. b) – Scree test showing eigenvalues for the y*y' matrix of the MDS: there is no obvious elbow in these values indicating that all global properties have a unique (if unequal) contribution to the scene representation.

References

Alvarez, G., & Oliva, A. (in press). The representation of simple ensemble visual features outside the focus of attention. Psychological Science.

Appelton, J. (1975). The Experience of Landscape. London: Wiley.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. Psychological Science, 12, 157- 162.

Ashby, F., & Lee, W. (1991). Predicting similarity and categorization from identification. Journal of Experimental Psychology: General, 120(2), 150-172.

Bar, M. (2004). Visual objects in context. Nature Reviews: Neuroscience, 5, 617-629.

Biederman, I. (1972). Perceiving real-world scenes. Science, 177, 77–80.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.) Perceptual Organization. pp. 213-263. Hillsdale, New Jersey: Lawrence Erlbaum.

Biederman, I. (1987). Recognition by components: A theory of human image understanding. Psychological Review, 94(2), 115-147.

Biederman, I., Rabinowitz, J.C., Glass, A.L., & Stacy, E.W. (1974). On the information extracted from a glance at a scene. Journal of Experimental Psychology, 103, 597-600.

Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. Cognitive Psychology, 14, 143-177.

Brainard, D.H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 443-446.

Bülthoff, H., Edelman, S., & Tarr M. (1995). How are three-dimensional objects represented in the brain? Cerebral Cortex, 3, 247-260

Chaumon, M., Drouet, V., & Tallon-Baudry, C. (in press). Unconscious associative memory affects visual processing before 100 ms. Journal of Vision.

Chen, L. (2005). The topological approach to perceptual organization. Visual Cognition, 12(4), 553-637.

Chong S. C., & Treisman, A. (2003). Representation of statistical properties. Vision Research, 43, 393-404.

Chong, S., & Treisman, A. (2005). Statistical processing: computing the average size in perceptual groups. Vision Research, 45(7), 891-900.

Chubb, C., Man, J., Bindmanm D., & Sperling, G. (2007) The three dimensions of human visual sensitivity to first-order contrast statistics. Vision Research, 47(17), 2237-2248.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive Psychology, 36, 28-71.

Cutting, J. (2002). Representing motion in a static image: constraints and parallels in art, science and popular culture. Perception, 31(10), 1165-1193.

De Graef, P., Christaens, D., & d'Ydewalle, G. (1990) Perceptual effects of scene context on object identification. Psychological Research, 52, 317-329.

Delorme, A., & Thorpe, S. (2003). SpikeNET: an event-driven simulation package for modeling large networks of spiking neurons. Network: Computation in Neural Systems, 14, 613-627.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local environment. Nature, 392, 598-601.

Evans, K., & Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free? Journal of Experimental Psychology: Human Perception and Performance, 31(6), 1476-1492.

Fei Fei, L., & Perona, P. (2005). A Bayesian Hierarchical model for learning natural scene categories. IEEE Proceedings in Computer Vision and Pattern Recognition, 2, 524-531.

Fei Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? Journal of Vision, 7(1), 1-29.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Vol. 2, II-264-II-271.

Freyd, J. (1983). The mental representation of movement when static stimuli are viewed. Perception and Psychophysics, 33, 575-581.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory of scene gist. Journal of Experimental Psychology: General, 108, 316-355.

Gibson, J.J. (1958). Visually controlled locomotion and visual orientation in animals. British Journal of Psychology, 49(3), 182-194.

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

Goh, J.O.S., Siong, S.C., Park, D., Gutchess, A., Hebrank, A., & Chee, M.W.L. (2004). Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. Journal of Neuroscience, 24, 10223-10228.

Gordon, R. (2004). Attentional allocation during the perception of scenes. Journal of Experimental Psychology: Human Perception and Performance, 30(4), 760-777.

Gosselin, F., & Schyns, P. (2001). Why do we SLIP to the basic level? Computational constraints and their implementation. Psychological Review, 108(4), 735-758.

Greene, M. R., & Oliva, A. (2005). Better to run than hide: The time course of naturalistic scene decisions, Journal of Vision., 5, 70a.

Greene, M.R., & Oliva, A. (2006). Natural Scene Categorization from Conjunctions of Ecological Global Properties. Proceedings of the 28th Annual Conference of the Cognitive Science Society, Vancouver, Canada, (pp. 291-296).

Greene, M.R., & Oliva, A. (in preparation, a) Different diagnostic image information for natural and indoor scenes.

Greene, M.R., & Oliva, A. (in preparation, b). High-level aftereffects to natural scenes.

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

Henderson, J. & Hollingworth, A. (2003). Global transsaccadic change blindness during scene perception. Psychological Science 14(5), 493-497.

Hollingworth, A., & Henderson, J. (1998) Does consistent scene context facilitate object perception? Journal of Experimental Psychology: General, 127(4), 398-415.

Joubert, O., Fize, D., Rousselet, G., & Fabre-Thorpe, M. (2005). Categorization of natural scenes: global context is extracted as fast as objects. Perception, 34s, 140.

Joubert, O., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. Vision Research, 47: 3286-3297.

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. Visual Cognition, 15(1), 12-19.

Kaplan, S. (1992). Environmental Preference in a Knowledge-Seeking, Knowledge-Using Organism. In J. H. Barkow, L. Cosmides, and J. Tooby (Eds.) The Adaptive Mind. New York: Oxford University Press, 535-552.

Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: a critical review. Psychological Bulletin, 112, 24-38.

Kourtzi, Z., & Kanwisher, N. (2000). Activation of human MT/MST by static images with implied motion. Journal of Cognitive Neuroscience, 12(1), 48-55.

Leopold, D., O'Toole, A., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. Nature Neuroscience, 4, 89-94.

Maljkovic, V. & Martini, P. (2005). Short-term memory for scenes with affective content. Journal of Vision, 5(3), 215-229.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York: Henry Holt and Co., Inc.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. (2005). The use of visual information in natural scenes. Visual Cognition, 12, 938-953.

Merilaita, S. (2003). Visual background complexity facilitates the evolution of camouflage. Evolution, 57(6), 1248-1254.

Motoyoshi, I., Nishida, S., Sharan, L, & Adelson, E. (2007). Image statistics and the perception of surface qualities. Nature, 447, 206-209.

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

Noesselt, T., Hillyard, S., Woldorff, M., Schoenfeld, A., Hagner, T., Jäncke, L., Tempelmann, C., Hinrichs, H. & Heinze, H. (2002). Delayed Striate Cortical Activation during Spatial Attention. Neuron, 35(3), 575-587.

Oliva, A., & Schyns, P. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive Psychology, 34, 72-107.

Oliva, A., & Schyns, P. (2000). Diagnostic colors mediate scene recognition. Cognitive Psychology, 41, 176-210.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. International Journal of Computer Vision, 42, 145-175.

Oliva, A., & Torralba, A. (2002). Scene-centered description from spatial envelope properties. Proc. 2nd International Workshop on Biologically Motivated Computer Vision, Eds: H. Bulthoff, S.W. Lee, T. Poggio, & C. Wallraven. Springer-Verlag, Tuebingen, Germany (pp.263-272).

Oliva, A., Mack, M.L., Shrestha, M., & Peeper, A. (2004). Identifying the Perceptual Dimensions of Visual Complexityof Scenes. Proc. 26th Annual Meeting of the Cognitive Science Society. Chicago.

Oliva, A. & Torralba, A. (2006). Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research: Visual Perception, 155, 23-36.

Palmer, S.E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. Norman & D. Rumelhart (Eds) *Explorations in Cognition* (p. 279-307) Hillsdale, NJ: Erlbaum.

Parkes, L., Lund, J., Angelucci, A., Solomon, J., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. Nature Neuroscience, 4, 739-744.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.

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

Potter, M.C., Staub, A., & O' Connor, D.H. (2004). Pictorial and Conceptual Representation of Glimpsed Pictures. Journal of Experimental Psychology: Human Perception and Performance, 30, 478-489.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. Behavior and Brain Sciences, 22, 341–423.

Ramachandran, V., Tyler, C., Gregory, R., Rogers-Ramachandran, D., Duessing, S., Pillsbury, C., & Ramachandran, C. (1996). Rapid adaptive camouflage in tropical flounders. Nature, 379, 815-818.

Rensink, R.A. (2000). The dynamic perception of visual scenes. Visual Cognition, 7(1/3), 17-42.

Rensink, R. A. O'Regan, J. K. Clark, J. J. (1997). To See or Not to See: The Need for Attention to Perceive Changes in Scenes. Psychological Science, 8(5), 367-373.

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

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

Russell, B.C., Torralba, A., Murphy, K.P., & Freeman, W.T. (in press). LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision.

Rousselet, G. A. Joubert, O. R. Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? Visual Cognition, 12(6), 852-877.

Saiki, J., & Hummel, J. (1998) Connectedness and the integration of parts with relations in shape perception. Journal of Experimental Psychology: Human Perception and Performance, 24(1), 227-251.

Sanocki, T. (2003). Representation and perception of spatial layout. Cognitive Psychology, 47, 43-86.

Schmolesky, M. T. Wang, Y. Hanes, D. P. Thompson, K. G. Leutgeb, S. Schall, J. D. Leventhal, A. G. (1998). Signal Timing Across the Macaque Visual System. Journal of Neurophysiology, 79(6), 3272-3278.

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

Serre, T., Oliva, A., & Poggio, T. A. (2007). A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences, 104 (15), 6424-6429.

Simons, D. (2000). Current approaches to change blindness. Visual Cognition, 7(1), 1-15.

Steeves, J.K.E., Humphreys, G.K., Culham, J.C., Menon, R.S., Milner, A.D., & Goodale, M.A. (2004). Behavioral and neuroimaging evidence for a contribution of color and texture information to scene classification in a patient with Visual Form Agnosia. Journal of Cognitive Neuroscience, 16, 955-965.

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

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. IEEE Pattern Analysis and. Machine Intelligence, 24, 1226-1238.

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

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological Review, 113, 766-786.

Torralba, A., Fergus, R., & Freeman, W. (2007). Tiny Images. MIT-CSAIL Technical Report 2007-024.

Tversky, A. (1977). Features of similarity. Psychological Review, 84(4), 327-352.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. Cognitive Psychology, 15(1), 121-149.

Ullman, S. (1999) High-level vision: object recognition and visual cognition. Cambridge: MIT Press.

Vogel, J., Schwaninger, A., Wallraven, C., & Bülthoff, H. (2006). Categorization of natural scenes: global vs. local information. Symposium on Applied Perception in Graphics and Visualization APGV. 153, 33-40.

Vogel, J., & Schiele, B. (2007). Semantic scene modeling and retrieval for content-based image retrieval. International Journal of Computer Vision, 72(2), 133-157.

Wade, N. & Verstraten, F. (2005). Accommodating the past: a selective history of adaptation. In C. Clifford & G. Rhodes (Eds) Fitting the Mind to the World: Adaptation and after-effects in high-level vision. (p. 83-102) New York: Oxford University Press.

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

Warren, W., Kay, B., Zosh, W., Duchon, A., & Sahuc, S. (2001). Optic flow is used to control human walking. Nature Neuroscience, 4(2), 213-216.

Webster, M., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural face categories. Nature, 428(6982), 557-561.

Wolfe, J. (1998). Visual memory: what do you know about what you saw? Current Biology, 8, R303-R304.

Chapter 3: The time course of natural scene understanding

Published as:

Greene, M.R., & Oliva, A. (2009) The briefest of glances: the time course of natural scene understanding. *Psychological Science.* 20(4), 464-472.

Introduction

Catching meaning at a glance is a survival instinct, and a uniquely human talent that movie producers manipulate to their advantage when making trailers: by mixing snapshots of meaningful scenes in a rapid sequence, they can convey in a few seconds an evocative story from unrelated pictures of people, events and places. In the laboratory, now classic studies have shown that novel pictures can be identified in a 10Hz sequence, although they are quickly forgotten when new images come into view (Intraub, 1981; Potter, 1975; Potter & Levy, 1969). While several studies have investigated the availability of visual features over the course of a glance, here we investigate the early perceptual availability of a number of semantic scene tasks. What types of meaningful information can human observers perceive from the briefest glances at novel scene images?

A typical scene fixation of 275-300 ms (Henderson, 2003; Rayner, 1998) is often sufficient to understand the "gist" of an image, namely its semantic topic (e.g. "birthday party": Intraub, 1981; Potter, 1975; Tatler, Gilchrist & Risted, 2003). It takes slightly more exposure to recognize the smaller objects in the scene (Fei-Fei, Iyer, Koch & Perona, 2007), or to report their locations and relations (Evans & Treisman, 2005; Tatler et al, 2003).

There is also evidence that sophisticated scene analysis can be accomplished by observers after viewing a novel scene for a single monitor refresh (10-40ms) without masking. Observers are able to classify real-world scenes using tasks as diverse as detecting how pleasant a scene is (Kaplan, 1992), whether a scene is natural or urban

(Joubert, Rousselet, Fize & Fabre-Thorpe, 2007); determining the basic or superordinant level categories of a scene (Oliva & Schyns, 2000; Rousselet, Joubert, & Fabre-Thorpe, 2005), or determining the presence of a large object (Thorpe, Fize & Marlot, 1996; Van Rullen & Thorpe, 2001). While the extraordinarily high performances in these studies may be partially mediated by the persistence in iconic memory, high performances are seen on similar tasks using masking paradigms (Bacon-Mace, Mace, Fabre-Thorpe & Thorpe, 2005; Fei-Fei et al, 2007; Greene & Oliva, in press; Grill-Spector & Kanwisher, 2005; Maljkovic & Martini, 2005).

While many studies of natural scene understanding have focused on basic-level categorization or object identification, real world scenes contain a wealth of structural and functional information whose time course of perceptual availability has not yet been determined. For example, how *navigable* a place is, or what environments afford *concealment* are perceptual decisions with high survival value (Kaplan, 1992). Similarly, how scene surfaces extend in space and how they change over time may influence how observers would behave in the scene. Spatial layout properties such as the *mean depth* of an environment, or its *openness* also influence its affordances (Oliva & Torralba, 2001). One can run in an open field, but not a small and enclosed cave. Some materials of natural environments have a high *transience* (e.g. the scene changes very rapidly from one glance to the next, as a rushing waterfall or a windy sand-scape), whereas others surfaces such as cliff rocks have low transience, changing mostly in geological time. Similarly, material properties of surfaces, along with the interplay of atmospheric elements (e.g., water, wind, heat) give a place a particular physical *temperature*, another global property of the natural environment that strongly influences observers' behavior.

All of these properties (and certainly more) combine to provide an understanding of the scene, much like the recognition of a face's gender, race and emotion are part of a person's identity, or how the recognition of an object depends on its shape, material, or pose.

In the present study, we establish perceptual benchmarks of early scene understanding by estimating the image exposure thresholds needed to perform two types of tasks: A basic-level scene categorization task performed in several blocks (whether an image is an ocean, a mountain, etc.) and a global property categorization task, where observers classified several spatial and functional properties of the scene image, also performed in different blocks (i.e. is the scene a hot place? Is it a large environment?). There are several possible predictions of the results based on different theories from the literature. Prototype theorists might predict that the basic-level categories should be available first, as this level is privileged in object naming experiments (e.g. Rosch, 1978). However, formal and experimental work has shown that global property information is highly useful for basic-level scene categorization (Greene & Oliva, in press; Oliva & Torralba, 2001), which would predict an early advantage for global properties. However, recent work examining the perceptual availability of object information at different levels of categorization has shown that while subordinant-level categorizations take more image exposure than basic-level categorizations, there was no presentation time difference between knowing that an object is present (versus noise) and knowing what it is at the basic level (Grill-Spector & Kanwisher, 2005), so there may be no substantial threshold differences between tasks.

Experiment

In psychophysics, staircase methods have been successful in efficiently determining human perceptual abilities (Klein, 2001). Here, we employ a presentation duration threshold paradigm to determine perceptual benchmarks on both global property and basic-level categorization tasks.

Method

Participants

20 participants (8 males, age 18-35) completed the psychophysical threshold experiment. They all had normal or corrected-to-normal vision and provided informed written content. They received $10 for the one hour study.

Stimuli

A total of 548 full-color photographs of natural landscapes were used in this experiment (see Figure 1). Images were 256x256 pixels in size and were selected from a large scene database (Greene & Oliva, in press; Oliva & Torralba, 2001).

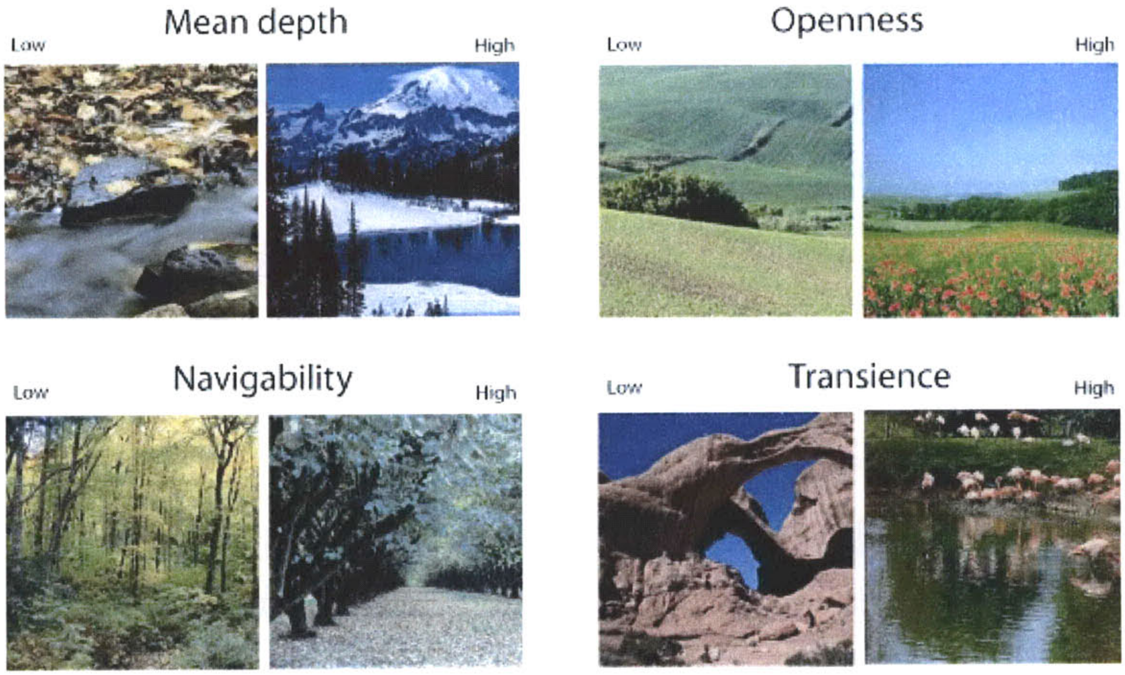


Figure 1: Example images from low and high poles of four global properties.

To compare natural image tasks, it is necessary to have normative rankings on the basic-level category and global property status of all images.

For the basic-level category classification blocks, we used prototypical scenes from 7 natural landscape categories (*desert, field, forest, lake, mountain, ocean* and *river*). Prototypicality of scenes' basic-level categories was assessed in a previous study (Greene & Oliva, in press) as follows: 10 naïve observers ranked 500 scenes on different basic-level category labels using a 1 (atypical) to 5 (highly prototypical) scale. At least 25 images per basic-level category with a mean rank of 4 or higher were selected. Additional exemplars were added for each category by visual similarity matching between the ranked prototypes and images from a database of ~10,000 natural landscapes. For each basic-level category block, 50 images were from a single target category (forest, for example) and 50 images were randomly selected from all other categories (constrained to have roughly equal numbers of each other category).

For the global property blocks, we used images that had been ranked as poles in one of 7 global properties (*concealment*, *mean depth*, *naturalness*, *navigability*, *openness transience* and *temperature*, see Greene & Oliva, in press, and Table 1 for descriptions). The same collection of 500 natural scene images were ranked along each of the global properties (excepting *naturalness*) using a hierarchical grouping task: at least 10 observers organized trials of 100 images at a time from lowest to greatest degree of a property (from the most close-up to farthest view when ranking *mean depth*, for example). Images whose ranks were within the first (<25%) or last quartiles (>75%) of the ranking range were considered typical poles for that global property and were used in the current experiment. Images for *naturalness* consisted of images sampled from this pool of natural images as well as various urban distractor images. For each global property block, 50 images from the high global property pole served as targets (high *openness*, or large *depth*, for example), and 50 images from the low pole served as distractors (e.g. closed or small depth). A description of the 7 global properties, as described to participants, is listed in Table 1.

Global Property	Target Description	Non-target description
Concealment	Scene contains many accessible hiding spots, and there may be hidden objects in scene.	If standing in the scene, one would be easily seen.
Mean depth	Scene takes up kilometers of space.	Scene takes up less than a few meters of space.
Naturalness	Scene is a natural environment.	Scene is a man-made, urban environment.
Navigability	Scene contains a very obvious path that is free of obstacles.	Scene contains many obstacles or difficult terrain.
Openness	Scene has a clear horizon line with few obstacles.	Scene is closed with no discernable horizon line.
Temperature	Scene environment depicted is a hot place.	Scene environment depicted is a cold place.
Transience	One would see motion in a video made from this scene.	Scene is not changing, except for patterns of daylight.

Table 1: Description of global property target and non-target images as described to participants in the experiment.

As far as possible, test images for both the category and global property tasks were drawn from the same population of natural landscape pictures. About half of all images served as both targets and distractors for different blocks. This helps to ensure that image-level differences are balanced across the experiment.

To produce reliable perceptual benchmarks, it is necessary to effectively limit additional sensory processing following image presentation. To this end, we used a dynamic masking paradigm (Bacon-Mace et al, 2005) consisting of a rapid serial visual presentation sequence of mask images. The use of multiple mask images minimizes visual feature interactions between target images and masks, ensuring a more complete masking of image features.

Mask images (Figure 2) were synthesized images created from the same database of natural images, using a texture synthesis algorithm designed by Portilla & Simoncelli (2000). We used the Matlab code provided on their web site enhanced to include the color distribution of the model input image. Examples of masks are shown in Figure 2. The texture synthesis algorithm uses a natural image as input, and then extracts a collection of statistics from multi-scale, multi-orientation filter outputs applied onto the image and finally, coerces noise to have the same statistics. Importantly, this method creates a non-meaningful image that conserves marginal and first-order statistics as well as higher-order statistics (cross-scale phase statistics, magnitude correlation and autocorrelation) while discarding object and spatial layout information. Additionally, a t-test performed on the power spectrum slopes for various orientations between the group of natural images and the group of masks was not significant (p*rep*=0.76).

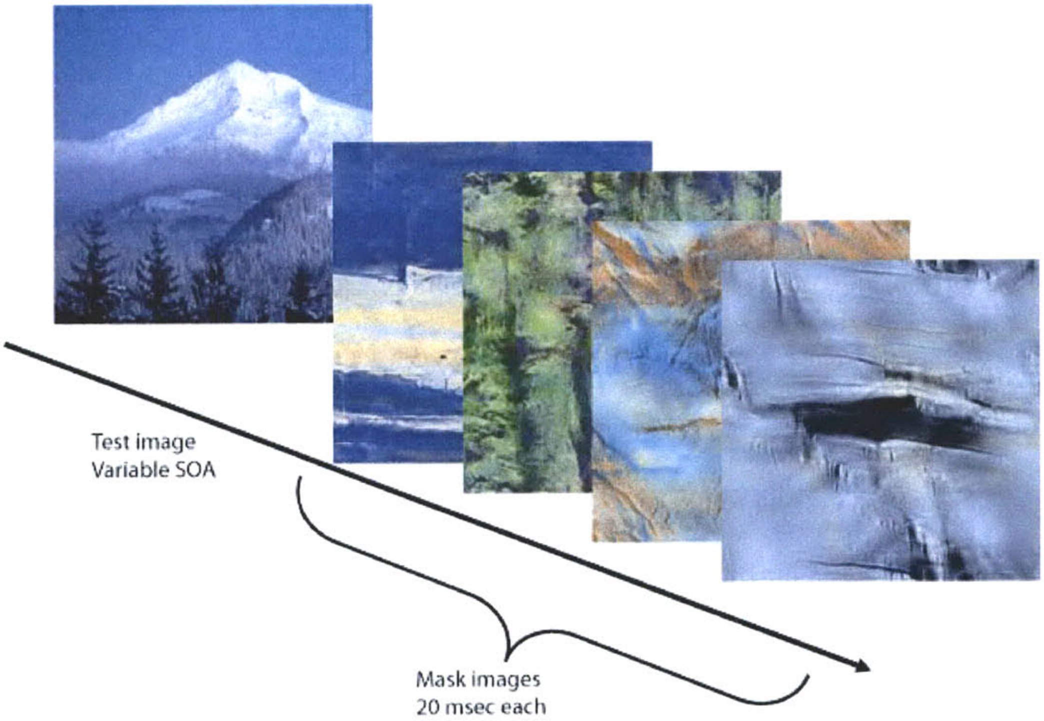


Figure 2: Schematic of experimental trial. The presentation duration of each test image was staircased using a linear 3-up-1-down procedure, with the first trial of each block presented for 50ms stimulus onset asynchrony (SOA). Test images were dynamically masked using four colored textures.

Design and Procedure

Participants sat in a dark room about 40cm away from a 21 inch CRT monitor (100Hz refresh rate). Stimuli on the screen subtended 7 deg. x 7 deg. of visual angle. Each participant completed 14 blocks of 100 images each: 7 category blocks and 7 global property blocks. The order of blocks was randomized and counterbalanced across participants. For each block, participants performed a yes-no forced choice task, and were instructed to respond as quickly and accurately as possible whether the image briefly shown was of the target (category or global property pole).

During each block, a linear 3-up-1-down staircase was employed. The first image in each block was shown for 50ms followed by the dynamic mask. Subsequent presentation times of trials in that block were determined by the accuracy of the observer's previous response, increasing by 10ms (to a ceiling of 200ms) if the response was incorrect and decreasing by 30ms (to a floor of 10ms) for a correct response. In this way, performance converges at 75% correct (Kaernbach, 1990).

At the beginning of each experimental block, an instruction page appeared on the screen, describing the task (detect a basic level category or a pole of a property, see Table 1) and giving a pictorial example of a target and a non-target. Figure 2 shows a pictorial representation of a trial. Each trial commenced with a fixation point for 250ms, followed by the target image for a variable presentation time (10-200ms staircased). Target images were immediately followed by a sequence of four randomly drawn mask images, presented for 20ms each, for a total of 80ms. Participants were then to respond to the target status of the image as quickly and accurately as possible. Visual feedback was provided for incorrectly classified images (the word "Error" displayed for 300ms following the response). Participants were first given a practice block of 20 trials to get used to the staircase procedure. The task for the practice block was "indoor vs. outdoor" which was not used in the main experiment. This experiment was run using Matlab and Psychophysics toolbox (Brainard, 1997; Pelli, 1997).

Results

For all blocks, the image presentation threshold was the presentation duration required for a participant to achieve 75% accuracy on the task. For some participants, not all blocks yielded a stable threshold. Due to the adaptive nature of the stair casing algorithm, very poor performance at the beginning of the block could lead to a considerable number of trials spent at 200ms of image duration (the ceiling duration) such that final threshold calculations were artificially high. For all results reported, we exclude data where more than 10% of trials were spent at the maximum duration of 200ms. Altogether these trials constituted only 5% of the data, and were evenly distributed between global property and basic-level category blocks (t(13)<1). Below, we examine two processing benchmarks: (1) an upper bound of the exposure duration necessary to perform a categorization block, given by the maximum image duration seen by each participant during each block and (2) the 75% correct threshold duration to compare time needed for equivalent performance across blocks.

To ensure equal task difficulties, we compared the maximum image exposure needed by each participant in each block. As image presentation times were controlled adaptively in the staircase procedure, the longest presentation time seen by a participant corresponds to the duration where no classification errors were made (recall that errors resulted in increased subsequent presentation times). If global property and category tasks are of comparable difficulty, we would expect them to have similar maximum duration values. Indeed, the mean maximum duration for the global property task was 93ms and the mean for the category task was 91ms (t(19)<1, see Table 2). This result indicates that both tasks were of similar difficulty.

In order to reliably estimate the 75% correct presentation time thresholds, we employed two methods: (1) taking the mode image duration seen by each participant and (2) fitting a psychometric function (Weibull) to the accuracy data for each presentation time, and solving for the threshold. Reported thresholds in Table 2 are the average of the two estimates.

A classic method for estimating thresholds from up-down staircases such as ours is to take the mode stimulus value shown to a participant (Cornsweet, 1962; Levitt, 1971). The logic here is simple: by moving the presentation duration 30ms shorter for a correct response on a previous trial, and moving 10ms longer for an incorrect response, the participants will, over the course of the block converge on 75% correct performance (Kaernbach, 1990), viewing more trials around the perceptual threshold than above or below it.

As estimation with the mode is a rather coarse method, we also estimated thresholds from the psychometric function for each participant and each block. Here, a Weibull function was fit to the performance data (proportion correct) for each presentation time viewed using the maximum likelihood procedure. This function typically provides very good fits to psychometric data (Klein, 2001). To illustrate, Figure 3 shows the Weibull fit and a histogram of presentation times viewed by one participant for a global property block and a basic-level category block.

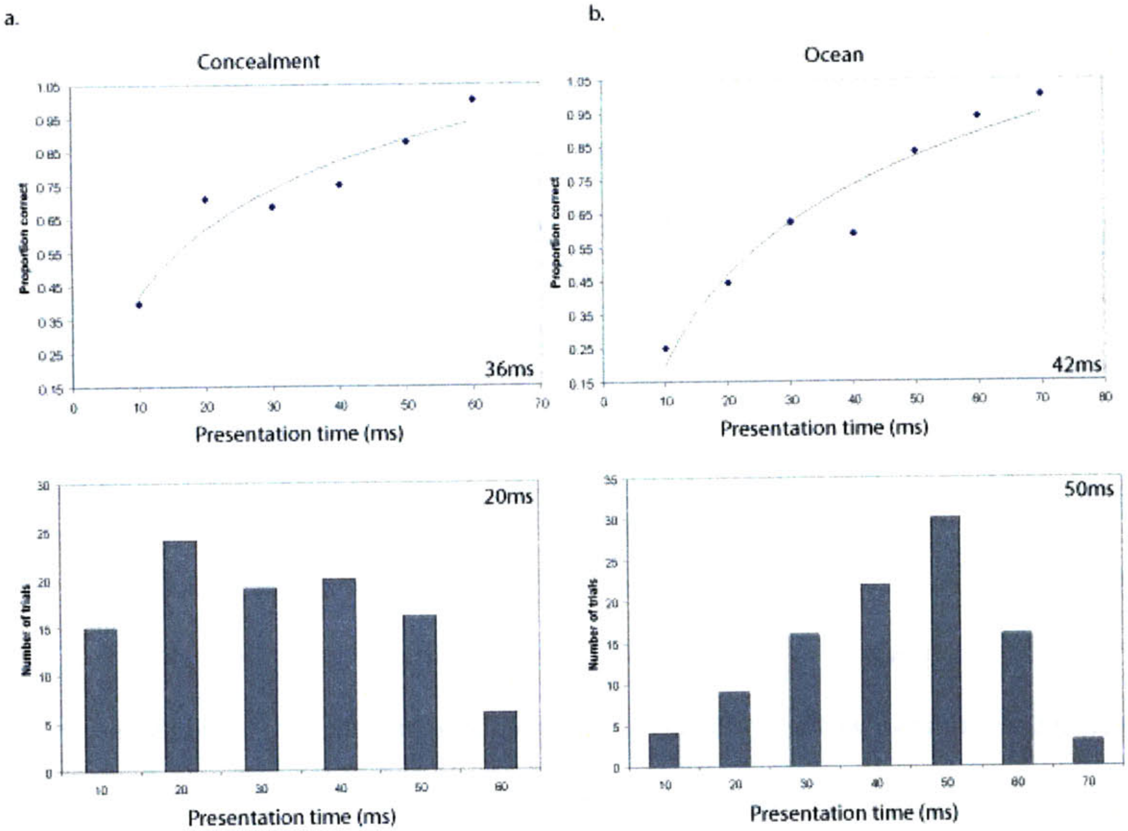


Figure 3: Example of threshold computation for example participant for (a) a global property block (concealment) and (b) a basic-level category block (ocean). Top row shows Weibull fits with thresholds, and bottom row shows histograms of presentation times viewed, where mode indicates thresholds. For all data reported here, the 75% presentation time threshold refers to the mean of these two values for each participant.

We found that the presentation time thresholds for all 14 categorization blocks were remarkably short (see Table 2): all were well under 100ms, and ranged from 19ms (naturalness) to 67ms (river).

	75% threshold	**Asymptote**
Concealment	**35** (2.7)	**97** (7.9)
Mean depth	**26** (2.8)	**75** (4.9)
Naturalness	**19** (1.9)	**63** (4.9)
Navigability	**36** (4.5)	**120** (9.2)
Openness	**47** (4.6)	**119** (9.5)
Temperature	**29** (2.4)	**119** (9.5)
Transience	**45** (4.0)	**123** (8.8)
Mean (st. dev.)	34 (10)	102 (24)

	75% threshold	Asymptote
Desert	**47** (4.7)	**93** (7.2)
Field	**55** (4.6)	**95** (7.3)
Forest	**30** (3.4)	**78** (6.6)
Lake	**51** (3.7)	**100** (7.1)
Mountain	**46** (3.3)	**95** (6.2)
Ocean	**55** (3.9)	**105** (6.5)
River	**67** (5.1)	**113** (6.1)
Mean (st. dev.)	50 (11)	97 (11)

Table 2: Presentation time threshold values (s.e.m. in parentheses) for the 7 global properties blocks (top table) and the 7 basic-level category blocks (bottom table). While global property blocks had lower average thresholds than basic-level category blocks, both reached asymptote performance at similar presentation times. While global property blocks had, on average, a smaller variance of thresholds between participants compared with category tasks (t(13)=-1.85, p*rep* = 0.83), there was larger variance in performance between the global property tasks, suggesting that these properties are less homogenous as a set than the basic-level categories.

We compared the threshold values for the global property blocks to the threshold values for the basic-level category blocks and found that the mean global property presentation time thresholds (34 ms) were significantly lower than the category thresholds (50 ms) (t(19) = -7.94, p*rep* = 0.99 for average thresholds; t(19) = 7.38, p*rep* > 0.99 for Weibull; t(19) = 3.51, p*rep* = 0.98 for mode). It is of note that to compare any tasks, it is necessary to ensure that there were equivalent distractor images. In the limit, a scene distractor with one pixel difference from the target would produce extremely large presentation time thresholds (if observers could perform the task at all). On the other hand, distinguishing scene targets from white noise distractors should result in ceiling performance. In our tasks, distractors were always prototypically different from the target image. For the global property blocks, this means that the distractors represented the opposite pole of the queried property and that both targets and distractors came from several basic-level categories. For the basic-level category blocks, this means that distractors were prototypes of a variety of other scene categories, and were chosen to show the greatest variety of category prototypes. In this way, targets and distractors were chosen, as best as possible, to vary only in the attribute being tested. Recall that global property and basic-level category tasks reached ceiling performance at similar presentation durations, indicating the equivalence of the distractor sets at longer presentation times.

Figure 4a shows the distributions of participants' presentation duration thresholds for global property blocks and basic-level category blocks. As shown in Figure 4b, the distributions of participants' thresholds in basic-level category blocks are rather homogenous in terms of both means and variances. In contrast, the distributions of thresholds on global property blocks (Figure 4c) are more heterogenous, some coming very early and others more closely resembling the category thresholds.

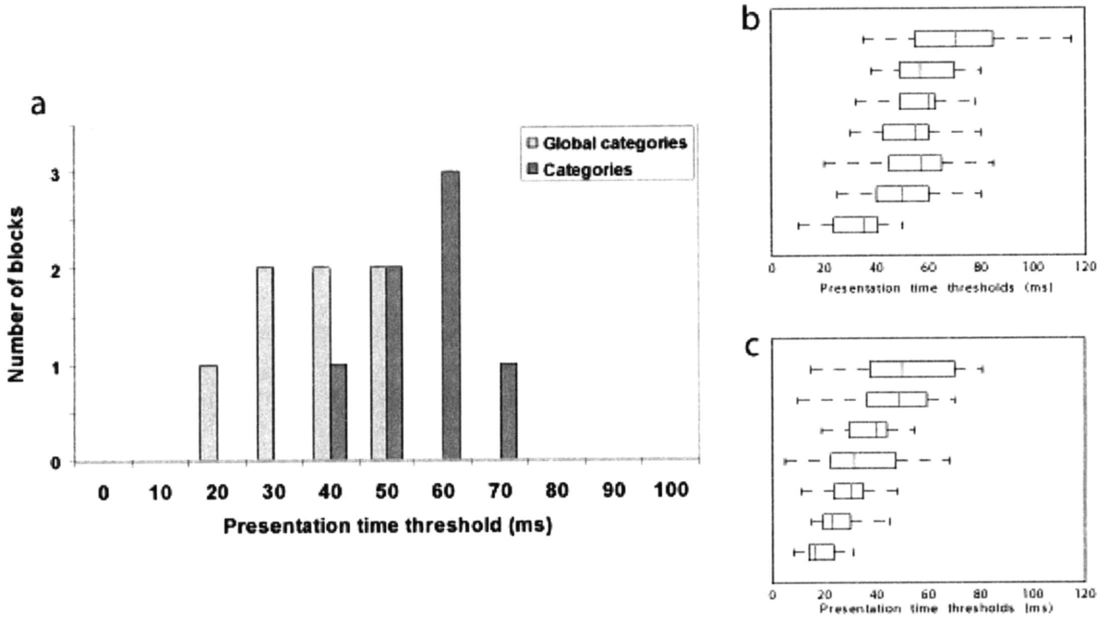


Figure 4: (a) Shows the distributions of observers' presentation duration thresholds for the global property tasks and the basic-level category tasks. (b) Shows the distribution for each basic-level category block. (c) Shows the distribution of each global property block. We calculated a 95% confidence interval around the global property and basic-level category means. We found that "forest" had a significantly faster threshold than other basic-level category blocks while "openness" and "transience" had significantly slower thresholds than other global property blocks.

Discussion

A large amount of meaningful information can be gleaned from a single glance at a scene (Bacon-Mace et al, 2005; Biederman, Rabinowitz, Glass & Stacy, 1974; Castelhano & Henderson, 2008; Fei-Fei et al, 2007; Grill-Spector & Kanwisher, 2005; Joubert et al, 2007; Maljkovic & Martini, 2005; Oliva & Schyns, 2000; Potter & Levy, 1969; Schyns & Oliva, 1994; Thorpe et al, 1996; Walker-Renninger & Malik, 2004 and many others), but our study is the first to establish perceptual benchmarks comparing the types of meaningful information that can be perceived during very early perceptual processing.

What meaningful perceptual and conceptual information can be understood from extraordinarily brief glances at a novel scene? Here, we provide insight into this question by comparing the shortest image exposures required for participants to achieve equivalent performance (75% correct) on a number of naturalistic scene tasks. We found that these benchmarks ranged from 19ms to 67ms of image exposure, reaching asymptote between 60 to 120ms of exposure. Remarkably, the perception of global scene properties required, on average, a lower presentation duration than the perception of the scene's basic-level category. These results are related to other works in ultra-rapid scene perception (Joubert et al, 2007; Rousselet et al, 2005) that demonstrated that reaction times in a natural versus manmade task were faster than to a semantic classification (e.g. mountain, urban). Indeed, we also found that *naturalness* classification required the least image exposure (19ms).

Our results are complementary to other studies examining the accrual of image information over time (Fei-Fei et al, 2007; Intraub, 1981; Rayner et al., in press; Tatler et al, 2003). For instance, Rayner et al (in press) found that while the overall semantic topic of a scene was rapidly understood, being able to find an object within that scene (such as a broom in a warehouse image) took at least a 150ms fixation. Likewise, in Fei-Fei et al (2007), observers were presented with briefly masked pictures depicting various events and scenery (e.g. a soccer game, a busy hair salon, a choir, a dog playing fetch) and asked to describe in detail what they saw in the picture. They found that global scene information, such as whether the picture was outdoor or indoor, was perceived well above chance (50%) with less than 100ms of exposure. Although free report responses may also be confounded with inference (overestimation of what was seen due to the covariance with other perceived features and objects, see Brewer & Treyans, 1981), and may be biased towards reporting verbally describable information, this study conferred with other results from the literature (Biederman et al, 1974; Intraub, 1981; Oliva & Schyns, 2000; Potter, 1975; Tatler et al, 2003 among others) finding that as image exposure increases, observers are better able to fully perceive the details of an image.

In agreement with a global-to-local view of scene perception (Navon, 1977; Oliva & Torralba, 2001; see also Joubert et al, 2007; Schyns & Oliva, 1994 and others), we have shown that certain global visual information can be more easily gleaned from an image than even its basic-level category at the very early stages of visual analysis. This result suggests the intriguing possibility that there exists a time during early visual processing that is sufficient for an observer to know that a scene is a natural landscape or a large space, but is insufficient to know it is a mountain or a lake scene. Our result may be predicted by computational work showing that basic-level scene categories cluster along global property dimensions describing the spatial layout of the scene (the *Spatial*

Envelope Theory; Oliva & Torralba, 2001). Furthermore, for human observers performing a rapid basic-level scene categorization task, more false alarms are produced by distractors sharing global property similarities with the target category than those that do not (for example, more false alarms to *closed* images when the target category was *forest*, Greene & Oliva, 2009). The current results lend credence to the possibility that rapid scene categorization may be achieved through the perception of a few robust global scene properties.

In the current study, the range of presentation time thresholds over all tasks was large (19-67ms), but remained well below 100ms of exposure. There was also a large range of thresholds within both global property and basic-level category tasks (19-47ms and 30-67ms respectively). This suggests substantial diversity in the diagnostic image information used by observers to perform each task, and that these pieces of information may be processed with different time courses. Future work will involve uncovering the image features responsible for these remarkable performances. An intriguing possibility that is now emerging from studies in visual cognition is the idea that the brain may be able to rapidly evaluate robust statistical summaries of features and objects, such as the mean size of a set of shapes (Ariely, 2001; Chong & Treisman, 2005), the average orientation of a pattern (Parkes, Lund, Angelucci, Solomon & Morgan, 2001); the center of mass of a set of objects (Alvarez & Oliva, 2008) or even the average emotion of a set of faces (Haberman & Whitney, 2007), in an automatic fashion (Chong & Tresiman, 2005) and outside of the focus of attention (Alvarez & Oliva, 2008). Similarly, some tasks might be performed with less presentation time than others because the features that are diagnostic of this task are somewhat coded more efficiently. For instance, *naturalness* had the fastest threshold in our study and the fastest reaction time in Joubert et al (2007), and has been shown to be correlated with low-level features, distributed homogeneously over the image (Torralba & Oliva, 2003). Likewise, Walker-Renninger & Malik (2004) demonstrated that texture statistics provided good predictions of human scene categorization at very short presentation times. By abstracting away statistical homogeneities related to structural and functional properties of a scene, the human brain may be able to comprehend complex visual information in a very short time. Uncovering the benchmarks of visual processing at the image feature level will be a significant step forward in understanding the algorithms of human visual processing.

References

Alvarez, G., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19(4), 392-398.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12, 157- 162.

Bacon-Mace, N., Mace, M.J.M., Fabre-Thorpe, M., Thorpe, S.J. (2005). The time course of visual processing: backward masking and natural scene categorization. *Vision Research*, 45, 1459-1469.

Biederman, I., Rabinowitz, J.C., Glass, A.L., & Stacy, E.W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103, 597-600.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 443-446.

Brewer, W., & Treyans, J. (1981). Role of schemata in memory for places. *Cognitive Psychology*, 13(2), 207-230.

Castelhano, M.S. & Henderson, J.M. (2008) The influence of color on scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, 34(3), 660-675.

Chong, S., & Treisman, A. (2005). Statistical processing: computing the average size in perceptual groups. *Vision Research*, 45(7), 891-900.

Cornsweet, T. (1962) The staircase method in psychophysics. *American Journal of Psychology*, 75(3), 485-491.

Evans, K., & Treisman, A. (2005). Perception of objects in natural scenes; is it really attention-free? *Journal of Experimental Psychology: Human Perception and Performance*, 31, 6, 1476-1492

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), 1-29.

Greene, M.R., & Oliva, A. (in press). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognitive Psychology*.

Grill-Spector, K., Kanwisher, N. (2005).Visual recognition: as soon as you know it is there, you know what it is. *Psychological Science*, 16(2), 152-160.

Haberman, J. & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17, 751-753.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences, 7*, 498-504.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 604-610.

Joubert, O., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. *Vision Research*, 47, 3286-3297.

Kaernbach, C. (1990). A single-interval adjustment-matrix procedure for unbiased adaptive testing. *Journal of the Acoustical Society of America*, 88, 2645-2655.

Kaplan, S. (1992). Environmental Preference in a Knowledge-Seeking, Knowledge-Using Organism. In J. H. Barkow, L. Cosmides, and J. Tooby (Eds.) *The Adaptive Mind.* New York: Oxford University Press, 535-552.

Klein, S. (2001). Measuring, estimating and understanding the psychometric function: A commentary. *Perception and Psychophysics*, 63(8), 1421-1455.

Levitt, H. (1971) Transformed up-down methods in psychoacoustics. *Journal of the*

Acoustical Society of America, 49, 467-477.

Maljkovic, V., & Martini, P. (2005) Short-term memory for scenes with affective content. *Journal of Vision*, 5(3), 215-229.

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

Oliva, A., & Schyns, P. (2000) Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176-210.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42, 145-175.

Parkes, L., Lund, J., Angelucci, A., Solomon, J., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4, 739-744.

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

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1), 49-71.

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

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, 81(1), 10-15.

Rayner, K. (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.

Rayner, K., Smith, T., Malcom, G., & Henderson, J. (in press). Eye movements and visual encoding during scene perception. *Psychological Science*.

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

Rousselet, G. A. Joubert, O. R. Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, 12(6), 852-877.

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

Tatler, B., Gilchrist, I., & Risted, J. (2003) The time course of abstract visual representation. *Perception*, 32, 579-593.

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

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

Van Rullen, R., & Thorpe, S. (2001). The time course of visual processing: from early perception to decision making. *Journal of Cognitive Neuroscience*, 13(4), 454-461.

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

Chapter 3: High-level aftereffects to global properties

Published as: Greene, M.R., & Oliva, A. (under review) Adapting to Scene Space: High-Level Aftereffects to Global Scene Properties.

Introduction

Just as a brief glance at a face can give a wealth of information about the person's age, gender, race, mood and attractiveness, a brief glance at a scene provides the observer with equally rich and varied information (Intraub, 1981; Potter, 1975; Oliva & Schyns, 2000). This brief glance can provide knowledge about whether the scene is indoors or outdoors (Fei-Fei, Iyer, Koch, & Perona, 2007); if outdoors, whether it is natural or urban (Greene & Oliva, 2009b; Joubert, Rousselet, Fize & Fabre-Thorpe, 2007; Rousselet, Joubert & Fabre-Thorpe, 2005); if there is a clear path for navigation (Greene & Oliva, 2009b; Kaplan, 1992), and even a sense of the pleasantness of the environment (Kaplan, 1992).

In addition to rapid processing, behavioral and computational work has shown that certain *global scene properties* that represent the structure and function of a scene (such as openness, mean depth, and potential for navigation) are correlated with a scene's basic level scene category (Greene & Oliva, 2009a; Oliva & Torralba, 2001). In a recent study, Greene & Oliva (2009a) observed that human observers' errors in rapid scene categorization were better predicted by the similarity between target and distractor images in a global property space than by similarity in an object space. For example, given a brief glimpse of a scene (50 ms), observers were more likely to confuse *river* and *forest* scenes which have very similar spatial layout properties (for example, both tend to be enclosed and concealed environments with a relatively low potential for efficient navigation), than to confuse *forest* and *field* scenes which have very different spatial layout properties, even though they have similar objects (for example, fields are more open than typical forests, and have greater potential for concealment and navigation).

Computational work has shown that a system can categorize pictures of scenes, particularly outdoor environments, by using localized combinations of low-level features such as texture elements, spatial frequency, orientation and color, without the need to segment the objects that compose the scene (Fei-Fei & Perona, 2005; Oliva & Torralba, 2001; Torralba & Oliva, 2002, 2003; Vogel & Schiele, 2007; Walker-Renninger & Malik, 2004). Altogether, these results suggest a global, *scene-centered* view of scene understanding in which the meaning of a scene can be understood from the rapid computation of global scene properties representing aspects of scene structure and affordance.

A scene-centered framework of recognition predicts that the visual system should be continuously updated to structural and functional regularities that are useful for recognition and action and therefore prone to adaptation along these dimensions. Just as adaptation is observed in the relevant coding dimensions for faces such as emotion, gender and identity (Leopold, O'Toole, Vetter & Blanz, 2001; Webster, 2004), we would expect that the human visual system also adapts to scene properties that are relevant for scene analysis. Broadly, aftereffects are measured changes in the perceptual appearance of stimulus B after being adapted through prolonged exposure to stimulus A. The effects of adaptation are often *repulsive* in nature, meaning that stimulus B will appear less like its adaptor A. As it is generally thought that adaptation reflects strategies used by neural system for optimizing perceptual mechanisms (Attnaeve, 1964; Barlow, 1961), the adaptation method has been long employed in psychology to elucidate neural mechanisms of perception (see Clifford, Wenderoth & Spechor, 2000; Clifford, Webster, Stanley, Stocker, Kohn, Sharpee & Schwartz, 2007; Wade & Verstraten, 2005 and Webster, 1996 for reviews).

Indeed, adaptation has been observed for many different features coded by the visual system, from basic features such as color, motion, orientation and spatial frequency (Wade & Verstraten, 2005) to higher-level properties such as facial emotion, gender and identity (Leopold et al, 2001; Webster, 2004). Adaptation has also been shown to transfer between sensory modalities (Konkle, Wang, Hayward & Moore, 2009). Furthermore, adapting to low-level image features can modulate higher level perceptual judgments. For example, adapting to lines curved like a smile can modulate perceived face emotion (Xu, Dayan, Lipkin & Qian, 2008); adapting to subtle relationships between dots can alter the perceived gender of point-light walkers (Troje, Sadr, Geyer & Nakayama, 2006); adapting to textures with different skewness can change the perceived glossiness of surfaces (Motoyoshi, Nishida, Sharan & Adelson, 2007) and adapting to textures with different orientation content can alter the perceived naturalness of real-world scenes (Kaping, Tzvetanov & Treue, 2007). The converse is also true: adaptation to the direction of implied motion from static photographs of movement (a racecar driving, for example) creates a measurable motion aftereffect in a random dot coherence measure (Winawer, Huk & Boroditsky, 2008). While each of these examples illustrates how low-level features can alter high-level perception and categorization (and vice versa), it has not yet been shown that adaptation to complex natural inputs such as scenes can alter the perception of subsequently presented natural scenes.

The goal of this work is to determine whether global aspects of natural scene structure and affordance can produce aftereffects that alter the perception of subsequently presented natural scenes. Intuitively, experiences from our daily lives tell us that this might be the case. After spending a day spelunking, the world outside of the cave might appear much larger than it did before. Many of us have had the experience of leaving our familiar environments to go on vacation in another place that looks very different from our homes, such as leaving a spacious suburb in California to visit New York City. Upon returning home, the differences in spatial layout between the two places might seem exaggerated: exposure to the urban, crowded, vertical structure of Manhattan might make the back yard seem spacious and green. If our visual system efficiently codes spatial and affordance properties of natural scenes, then we would expect observers to be sensitive to small differences in these properties' magnitudes, producing aftereffects. Furthermore, if these same global properties are used by the visual system for rapid scene categorization, then adaptation to these properties should alter the speed and accuracy of human scene categorization abilities.

Greene & Oliva (2009a) proposed a set of global scene properties designed to reflect the natural variation in natural scene categories' spatial, surface and affordance properties (see also Appelton, 1975; Gibson, 1979, Kaplan, 1992 & Oliva & Torralba, 2001). Importantly, human observers are sensitive to these properties in rapid scene categorization tasks (Greene & Oliva, 2009a), making them good candidate properties for aftereffects.

In Experiment 1, we tested for perceptual aftereffects from adaptation to five global properties of natural scenes (openness, naturalness, mean depth, navigability and temperature, see Figure 1 for pictorial examples) using a novel rapid serial visual presentation (RSVP) adaptation paradigm. Experiments 2-4 explore the nature of these aftereffects using the *openness* of a scene's space as the case study. In Experiment 2, we ruled out the possibility that the aftereffects observed in Experiment 1 were inherited from adapting low-level (retinotopic) visual areas, and in Experiment 3 we ruled out the possibility that the aftereffects are due to a post-perceptual decision bias. Last, Experiment 4 tested the extent to which participants' adapted state to a global property might contribute to rapid scene categorization ability, suggesting a causal role for global property computation at an early stage of scene representation. Taken together, these results indicate that certain global properties of natural scenes are selectively adaptable, producing high-level aftereffects, and that such properties may be relevant for the rapid categorization of natural scenes.

Experiment 1: Aftereffects to Global Scene Properties

The goal of the first series of experiments was to determine if aftereffects could be obtained for a set of global scene properties in a novel rapid serial visual presentation (RSVP) adaptation paradigm. Here, we tested five global properties (openness, mean depth, naturalness, navigability and temperature) for aftereffects. In these experiments, we adapted participants to the extremities (or poles) of each global property dimension. Figure 1 shows examples of the poles of each of these global property dimensions. Each global property was tested in an independent experimental session. As the method and design details for all of these experiments was the same, we are presenting the five experiments as one.

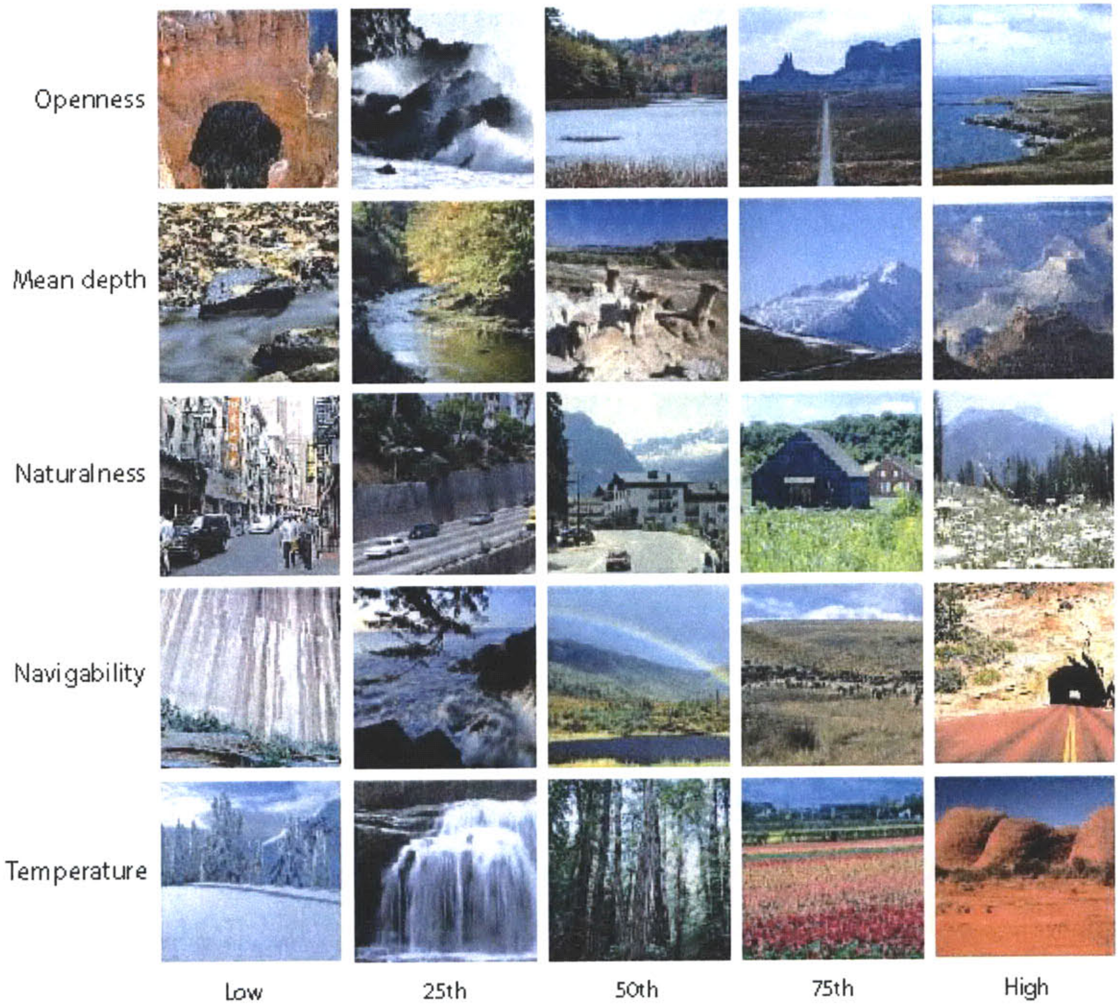


Figure 1: Example images illustrating the five global scene property used in Experiment 1. Images on the ends were used in the adaptation phase, and images from the 25th, 50th and 75th ranking percentiles were used as test images.

General Method

Materials

Scene images were full color, 256 x 256 pixels in size, and were chosen from a large laboratory database of real-world photographs that had been previously ranked along the dimensions of naturalness, openness, navigability, mean depth and temperature (Greene & Oliva, 2009a). To summarize, observers performed a hierarchical grouping task that organized groups of 100 images from lowest to greatest degree of each global property by making three binary groupings that produced eight groups of images. For example, observers organized the images from the most close-up to the farthest view for the case of mean depth, or from coldest to hottest places in the case of temperature. Detailed description of this ranking can be found in Greene & Oliva (2009a).

Adaptation and test images were chosen from these rankings. Adaptation images were chosen from the poles (or extremes) of the ranks, and test images were moderate along the ranks (see Figure 1 for pictorial examples). For each global scene property, three groups of 100 images were chosen. First, 200 images served as experimental adaptors, 100 from each pole of the property (for example, 100 images of *natural* environments and 100 *urban* environments in the case of naturalness). In all cases, these images were chosen to vary as much as possible in physical and semantic attributes other than the global property being tested. For example, in the case of *mean depth*, large depth images would consist of panoramic images from many natural image categories (fields, oceans, farmland, mountains, canyons, etc.) with various viewpoints, object density and lighting. The third group of 100 images served as a control adaptation condition, and represented all ranks along a given global property dimension. The test images consisted of 30 additional images for each global property that represented rank values from around the 25th, 50th and 75th ranking percentiles (see Figure 1 for examples).

All experiments were run using MATLAB and psychophysics toolbox (Brainard, 1997, Pelli, 1997). Experiments were displayed on a 21" CRT monitor with a 100 Hz refresh rate. Images subtended approximately 7 x 7 degrees of visual angle.

Participants

A total of 46 participants from the MIT community participated in at least one of the five experiments. Each global property was run as an independent experiment, so individual observers could participate in more than one experiment. Between 10 and 20 observers participated in each experiment. All were between 18-35 years old and had normal or corrected-to-normal vision. Participants provided informed consent and were paid $10/h for their time.

Design and procedure

Each of the five global properties was tested in an independent experimental session lasting approximately 45 minutes. Each experiment was a within subjects design in which participants were adapted to each pole of the global property and to the control set in three separate blocks. The order of the blocks was counterbalanced across participants.

A schema of the experimental procedure for a sample block is shown in Figure 2. Each experimental block consisted of two phases, an adaptation phase (Figure 2A) and a testing phase (Figure 2B). The adaptation phase was approximately five minutes long and consisted of displaying the 100 adaptor images eight times each in random order. Each image was shown for 100 ms with 100 ms blank between images. To keep focus on the image stream, participants were instructed to press the space bar when back-to-back image repeats were displayed. On average, there were seven repeats in the stream, appearing about every 80 seconds.

The testing phase consisted of 30 trials, and immediately followed the adaptation phase. Each trial commenced with 10 seconds of top-up adaptation were given in the form of a rapid serial visual presentation (RSVP) stream in which the 100 adaptor images were shown again for 100 ms each in random order. Participants were instructed to carefully watch and attend to the 10 second image stream. Following the top-up RSVP adaptation, stream there was a 500 ms blank, followed by the test image presented for 100 ms, and then masked by a 1/f noise mask for 80 ms. Following each test image, participants were instructed to respond as quickly and accurately as possible as to which pole of the global property the test image belonged. For example, in the *mean depth* experiment, participants would indicate if the test image was *large depth* or *small depth.* As test images were rated as ambiguous along the global property dimension tested, no performance feedback was given. The descriptions of the global properties as given to participants can be found in the Table 1.

Global property	High pole description	Low pole description
Mean depth	The scene takes up kilometers of space.	The scene takes up less than a few meters of space.
Naturalness	The scene is a natural environment.	The scene is a man-made, urban environment.
Navigability	The scene contains a very obvious path that is free of obstacles.	The scene contains many obstacles or difficult terrain.
Openness	The scene has a clear horizon line with few obstacles.	The scene is closed, with no discernible horizon line.
Temperature	The scene environment depicted is a hot place.	The scene environment depicted is a cold place.

Table 1: Description of global scene properties.

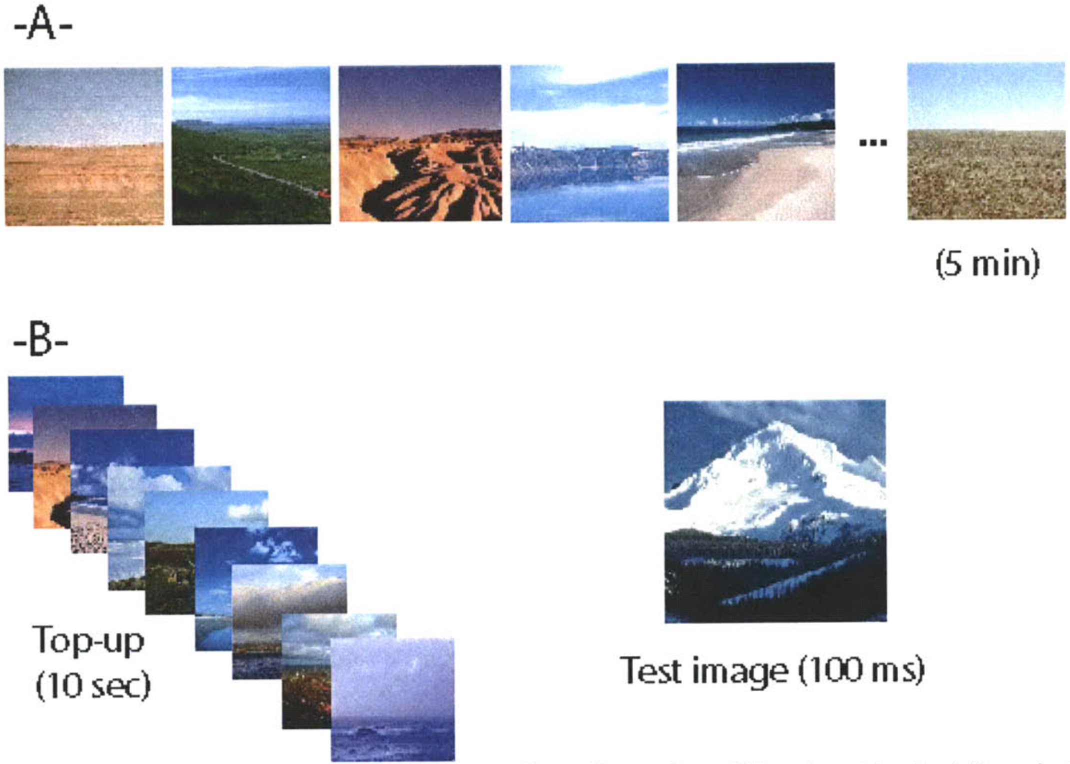


Figure 2: A schematic representation of the experimental procedure of Experiment 1. –A- A five minute adaptation phase in which participants viewed 800 adaptor images (100 images repeated 8 times each) while performing a one-back task. -B- Each trial of the test phase consisted in a 10 seconds top-up adaptation in the form of an RSVP stream, followed by a test image for 100 msec.

Results

As aftereffects are fleeting (Rhodes, Jeffery, Clifford & Leopold, 2007), speed was essential. At the test, trials with reaction times greater than two seconds were discarded from the analysis (the mean RT over the five experiments was around 760 ms). Participants whose mean reaction time was more than three standard deviations above the group mean were not included in the analysis (n=6). As each global property was tested independently, each was analyzed separately. As we did not have hypotheses about the relative magnitudes of the adaptation effects, no comparison between the properties is provided.

Figure 3 illustrates participants' responses in each of the five experiments. For each participant in each experiment, we computed the proportion of trials in which the participant classified test images as the high pole of the global property (i.e. open, natural, hot, large depth, and navigable) for each of the three groups of test images. Adaptation to each pole of a global property was compared against adaptation to the control stream, to establish a baseline for how the test images would be classified in our paradigm. As shown in Figure 3, participants' classifications of the *same* test scenes differed systematically with their adaptation condition. For example, adaptation to *open* images made moderately open images appear more *closed* than after viewing the control stream of images. Importantly, the same test images were perceived by the same observer as more open after adapting to closed images.

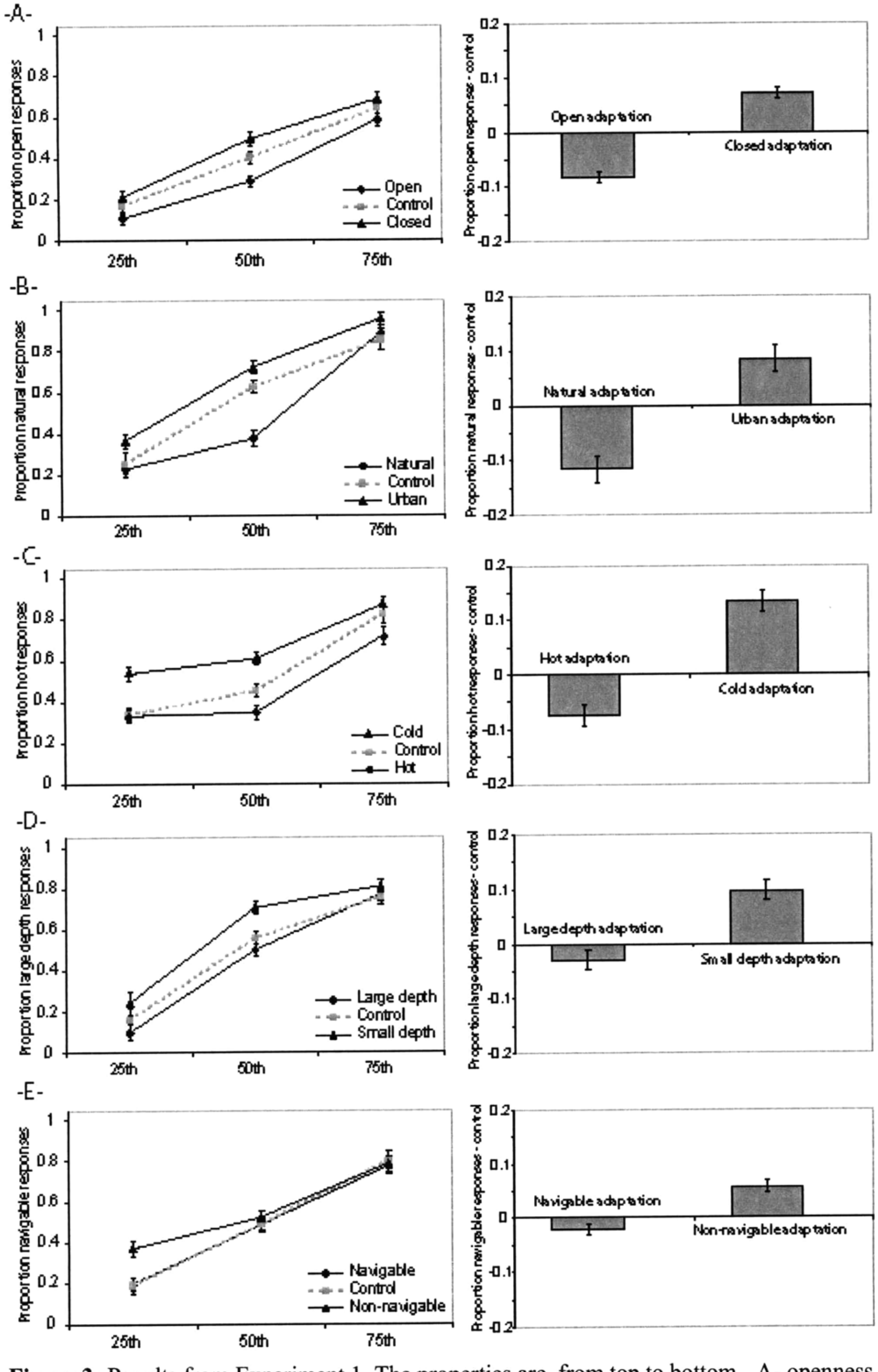


Figure 3: Results from Experiment 1. The properties are, from top to bottom, -A- openness, -B- naturalness, -C- Temperature, -D- Mean depth and –E- Navigability. Error bars correspond to +/- 1 within-subjects SEM (Loftus & Masson, 1994). Graphs in the left column show proportion of responses to the high pole of each global property for the three groups of test images over the three adaptation conditions. Graphs in the right column show the magnitude of the effect in each direction by showing the proportion of high pole responses for the two global property poles subtracted from responses to the control condition.

Repeated-measure ANOVA was performed on the average proportion of images classified as the high pole of the global property for each experimental session. There was a significant main effect of adaptation condition for *openness* (F(2,40)=19.51, p<0.001), *naturalness* (F(2,18)=10.8, p<0.001), *temperature* (F(2,30)=19.71, p<0.001), *mean depth* (F(2,30)=7.95, p<0.005) and *navigability* (F(2,26)=3.69, p<0.05). The mean magnitude of the aftereffects (the overall difference between adapting to one global property pole versus the other, and collapsing over the three groups of test images) was 21% for temperature, 20% for naturalness, 15% for openness, 13% for mean depth and 8% for navigability.

We next determined whether both poles of each global property showed significant adaptation. For each participant and for each adaptation condition, we collapsed over the three groups of test images and subtracted the proportion of responses to the high global pole of the global property from the proportion responses to the high pole from the control block. For each global property, we contrasted these with the null hypothesis that these numbers were zero, indicating the absence of aftereffects. Average magnitudes are shown in the right-hand column of Figure 3. For all properties except navigability, both global property poles were significantly different from zero (p<0.05).

Discussion

Here we have shown that several global scene properties related to scene spatial layout and function can produce aftereffects. Experiment 1 demonstrated robust aftereffects to four global properties (naturalness, openness, temperature and mean depth). The property *navigability* showed a weak and one-directional aftereffect as shown in Figure 3E.

To our knowledge, this is the first laboratory demonstration of aftereffects from prolonged viewing of natural scene images. However, we are all aware of similar effects in our daily lives, such as moving from a cramped airplane cabin into a spacious airport terminal. The global scene properties tested here are known to reflect a large amount of the variability existing between natural scene categories (Appelton, 1975; Baddeley, 1997; Gibson, 1979; Greene & Oliva, 2009a, 2009b; Joubert et al, 2007; Kaplan, 1992; Rousselet et al, 2005) and are informative dimensions describing differences between basic-level scene categories (Greene & Oliva, 2009a; Oliva & Torralba, 2001). Adaptation is generally seen as a functional mechanism used by the visual system to efficiently encode changes in the visual world (Attnaeve, 1954; Barlow, 1961). In this framework, the visual system can store an average (or prototype) value for a stimulus, and encode individual exemplars as differences from this prototype (Leopold et al, 2001). For environmental scenes, this prototype may reflect the mode of experienced scene properties. In other words, this prototype reflects the most common values of scene spatial layout and function that one has experienced. Finding stimulus dimensions that are prone to adaptation is informative for ascertaining neural mechanisms underlying perception as adaptation is believed to target neural populations underlying the processing of the adapted dimension. In other words, the existence of aftereffects for a particular stimulus dimension can be taken as evidence for neural populations representing that dimension.

An outstanding question is the extent to which the aftereffects observed in Experiment 1 are a result of adaptation of multiple low-level features, rather than adaptation of the global properties as single, high-level entities. Indeed, the global properties of *naturalness, openness* and *mean depth* are also well-correlated with low-level image features such as combinations of localized orientations and spatial frequencies (Oliva & Torralba, 2001; Torralba & Oliva, 2002). For example, a high degree of openness is correlated with low-spatial-frequency horizontal orientation in the vertical center of the image: a feature that corresponds with the horizon line of the scene, whereas a low degree of openness is correlated with more uniform texture throughout the image (Oliva & Torralba, 2001). Similarly, the judgment of how hot or how cold a place is (*temperature*) is related to the reflectance, color and material properties of scene surfaces, like the difference between desert sandstone and an iced-over river; and aftereffects have been observed to texture and material properties (Durgin & Huk, 1997, Motoyoshi et al, 2007). Therefore, it is possible that the aftereffects observed in Experiment 1 could be inherited from the low-level adaptation of visual features. We address the nature of global property aftereffects in Experiment 2.

Experiment 2: Translation Invariance of *Openness* Aftereffect

As robust aftereffects have been demonstrated for low-level features (for review, see Clifford et al, 2007), we need to address the extent to which the aftereffects observed in Experiment 1 are due to low-level adaptation of low-level features inherited from early visual areas.

A standard method for gaining insight into the processing level of aftereffects has been to test the translation invariance of these effects. As early visual areas have small receptive fields, adaptation of cells in these areas will not be invariant to a shift in location, while later visual areas show greater tolerance to this transformation (Gross, 1973; Ito, Tamura, Fujita & Tanaka, 1995). Melcher (2005) examined a variety of aftereffects, and found that the degree of spatial tolerance of the effects is related to the complexity of the stimulus: contrast adaptation had no spatial transfer, but faces had considerable transfer (c.f. Jiang, Blanz & O'Toole, 2006; Leopold et al, 2001; Rhodes et al, 2005; but see Afraz & Cavanagh, 2008). In Experiment 2, we tested the spatial tolerance of global scene property aftereffects, using the global property of *openness* as a test case.

A new group of participants were adapted to images centered five degrees of visual angle to the right or left of a central fixation. Aftereffects were probed in the opposite hemifield, five degrees away from fixation in the opposite hemifield from where adaptation occurred. If the aftereffects observed in Experiment 1 were inherited from adaptation of low-level visual features from early visual areas, then we would not expect to observe an aftereffect in Experiment 2. However, if the aftereffect is invariant to the hemifield transformation, then it suggests the existence of a high-level aftereffect.

Methods

Participants

10 new observers from the MIT community participated in Experiment 2. All were between 18-35 years old and had normal or corrected-to-normal vision. As eye fixation was monitored with an eye tracker, only participants without eye glasses were selected. Participants provided informed consent and were paid $10/h for their time.

Materials

The same set of images used for testing adaptation to *openness* in Experiment 1 was used here. Participants' right eye positions were monitored with an ETL 400 ISCAN table-mounted video-based eye tracking system sampling at 240 Hz. Participants sat at 75 cm from the display monitor and 65 cm from the eyetracking camera, with their head centered and stabilized in a headrest. The position of the right eye was tracked and viewing conditions were binocular.

Design and procedure

The design and procedure for Experiment 2 was identical to that of Experiment 1 except that the five minute adaptation phase and the top-up adaptation streams were presented at a location centered five degrees to one side of a central fixation point, while test images were centered five degrees on the other side. The side that was adapted was counterbalanced across participants. Images were approximately 5.3 x 5.3 degrees of visual angle in size, and there was no spatial overlap between adaptation and test locations. Eye position was monitored throughout the experiment, and trials in which the eyes moved more than one degree away from central fixation were discarded from analysis (this corresponds to two trials from one participant, none for all others).

Results

As in Experiment 1, for each participant, we computed the proportion of trials in which the participant classified test images as *open* for each of the three groups of test images. Also as in Experiment 1, trials with reaction times greater than two seconds were discarded from analysis. Repeated-measure ANOVA was performed on the average responses of each observer. There was a significant main effect of adaptation condition ($F(2,40)=8.83$, $p<0.05$) indicating that the openness aftereffect survived a ten degree spatial shift.

As in Experiment 1, we then tested whether the aftereffect was significant for both global property poles. Indeed, the *open* ($t(9)=3.12$, $p<0.05$) and *closed* ($t(9)=3.04$, $p<0.05$) poles showed significant aftereffects. The magnitude of the adaptation effect (the summed magnitude from each pole) was 14%, which was similar to the 15% magnitude observed in Experiment 1. This degree of spatial invariance is similar to the results obtained by the face adaptation literature (Afraz & Cavanagh, 2008; Jiang, Blanz & O'Toole, 2006; Leopold et al, 2001; Rhodes et al, 2003).

Discussion

Here we have shown that the *openness* aftereffect observed in Experiment 1 has strong position invariance, and is therefore unlikely to be solely due to the cumulative adaptation across multiple low-level features from early visual areas. This result suggests that what is being adapted is a higher-level representation of the degree of openness of a scene.

The current results show that there is substantial spatial transfer of aftereffects across space. Although we observed similar a similar magnitude of adaptation in this study, spatial transfer of face aftereffects typically find that that the magnitude of the effect is 50-70% of the magnitude of the aftereffect when tested in the adapted location.

Our current result suggests that the aftereffects observed in Experiments 1 and 2 are high-level in nature, and not simply inherited from adaptation of lower level features.

.

Experiment 3: Ruling out the Post-Perceptual Account

Experiments 1 and 2 found that participants were more likely to classify test images as more dissimilar to the global property pole that they were adapted to. In other words, scenes that were, for instance, moderately *natural* would appear more or less *natural* given the observer's adapted state. Given the difficulty in describing perceptual changes in a complex natural scene, we need to account for the possibility that in fact, the aftereffects observed in Experiments 1 and 2 could be explained by post-perceptual decision biases rather than perceptual aftereffects. In other words, were participants classifying these ambiguous scenes as less similar to the adapting images because the adapting images changed participants' decision boundaries between the global property poles?

As the participants' task in Experiments 1 and 2 was to determine the global property pole of a test image (*open* or *closed*, for example), it is possible that seeing many very open scenes would simply change the decision boundary between *open* and *closed* without producing a perceptual aftereffect. Our participants were not given feedback during the experiments, so they were not motivated to intentionally adopt a behavioral strategy that would shift the response curves. However, the results of Experiments 1 and 2 could also be explained by the decision boundary between global property poles shifting systematically with the adaptation condition. Ruling out this decision criterion is a potential problem for all experiments claiming high-level aftereffects, and while it is sometimes acknowledged (Troje et al, 2006), it has not been satisfactorily addressed.

In Experiment 3, we address the decision criterion issue by testing whether participants' adapted state to a global property pole would systematically influence an orthogonal basic-level scene categorization task. We reason that if adaptation to a global property (for example, the *openness* of an environment) changes observers' performance in a task that does not involve the judgment of that global property, then the change in classification performance is unlikely to be a result of a post-perceptual decision bias. We chose a basic-level categorization task for this purpose. There can be graded degrees of category membership in natural scene categories, and an image of a natural environment can lie between multiple basic-level categories. For example, a landscape image composed of trees, water and hills in the background has elements of forest, lake and mountain scene categories. In Experiment 3, we capitalize on the fact that there exists a continuum of environments between a *forest* prototype, which is typically an enclosed environment, and a *field* prototype which is typically open (see Figure 4 for examples). Therefore, if the classification changes observed in Experiments 1 and 2 were perceptual aftereffects, then adaptation to very open scenes should make an ambiguous image on the field-forest continuum look more like a forest, and adaptation to very closed scenes should make that image look more like a field. In Experiment 3, we used an adaptation method analogous to Experiment 1, where test images were exemplars ranked as lying between forest and field prototypes.

Methods

Participants

Twelve participants (9 new and 3 from Experiments 1 or 2) from the MIT community participated in this experiment. All were between 18-35 years old and had normal or corrected-to-normal vision. Participants provided informed consent and were paid $10/h for their time.

Materials

In this experiment, it is important that observers only adapt to the *openness* of environments. Therefore, we removed forest and field images from the adaptation streams, replacing them with images from other basic-level categories such as as ocean, canyon, desert, beach, etc.

Test images were chosen from a database of natural images previously ranked on their prototypicality in regards to various basic-level categories (Greene & Oliva, 2009a, Experiment 3). In this previous study, 10 observers ranked 500 images of natural landscapes in terms of how typical each image was for each of several basic-level scene category labels using a scale from 1 (atypical) to 5 (highly prototypical). For the current experiment, the test images consisted of 30 natural landscape images that had been ranked as partially prototypical for both *forest* and *field* categories. Analogous to Experiments 1 and 2, three groups of test images were chosen: 10 images that were ranked as more field than forest, 10 that were equally prototypical of field and forest and 10 that were more forest than field. Figure 4 shows example images along the ranked continuum between forest and field.

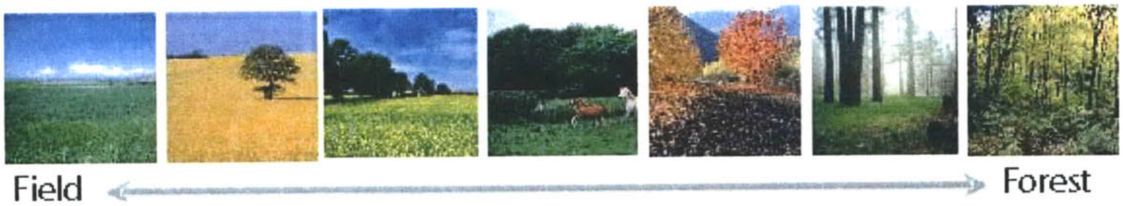


Figure 4: Examples of images ordered along the field-forest continuum. Environmental scenes, unlike most objects, can belong to more than one basic-level category. Experiment 3 tested images from the middle of the row, while Experiment 4 tested images from the ends.

Procedure

As in Experiments 1 and 2, each participant completed three experimental blocks that each contained two phases, an adaptation phase and a test phase. The adaptation phase of each block was identical to Experiment 1. Following the adaptation phase, participants completed a test phase that was identical to that of Experiment 1 except that the instructions were to classify test images as *forests* or *fields* as quickly and accurately as possible. As in Experiment 1, no performance feedback was given.

Results

Trials with reaction times greater than two seconds were discarded from analysis, and one participant with a mean reaction time of 3843 ms was not included in the analysis. As shown in Figure 5A, adaptation to openness modulated participants' basic-level classifications of natural scene images. After adapting to *open* images, participants were more likely to classify ambiguous test images as forests rather than fields. Conversely, after adapting to closed scenes, ambiguous test images were more likely to be categorized as fields (F(2,20)=17.87, p<0.001). The overall magnitude of the effect was 11% (see Figure 5A). While adapting to open scenes strongly modulated test image categorization as forest or field (t(10)=4.88, p<0.001), adaptation to closed images had only a marginally significant effect (t(10)=2.21, p~0.08).

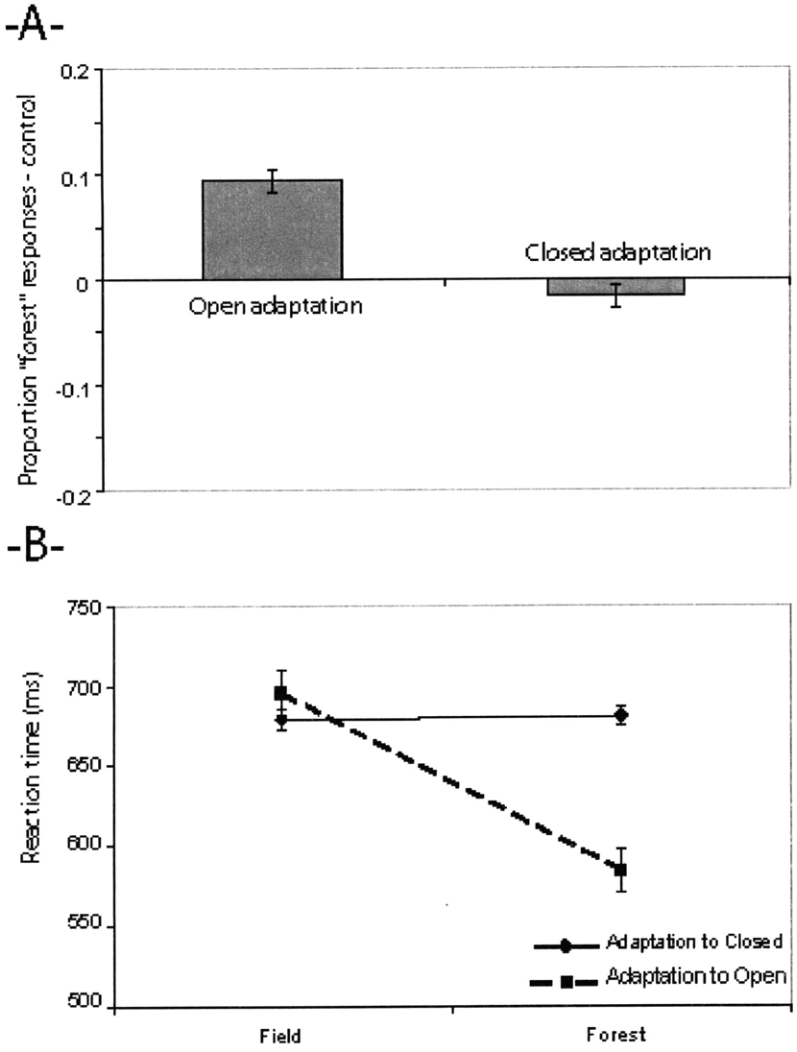


Figure 5: Results of basic-level categorization (field or forest) after adaptation to open or closed images. -A- Results of Experiment 3: Aftereffects to ambiguous images along the forest-field continuum: adapting to *open* scenes makes ambiguous images appear more like *forests*. –B- Results of Experiment 4: Reaction time to categorizing prototypical images of fields and forests, after adaptation to *open* and *closed* scenes.

Discussion

Here we observed that adaptation to the openness of natural environments can systematically shift the perception of a scene's basic-level category. For example, after adapting to very open environments, scenes such as ones in the middle of Figure 4 will look more like forests. However, these same images will look more like fields after adapting to closed environments. This result suggests that the aftereffects observed in Experiments 1 and 2 were not due to a change in decision criterion, but rather due to a perceptual aftereffect.

Determining whether classification changes following adaptation are due to perceptual or post-perceptual mechanisms is essential for all studies of high-level aftereffects, particularly when the adaptation dimension is continuous such as the global properties used here, or a continuously morphed face space (e.g. Jiang, Blanz & O'Toole, 2006; Leopold et al, 2001; Rhodes et al, 2005 and many others). While this issue is sometimes addressed in the literature as a potential weakness of high-level adaptation paradigms (Troje et al, 2006), or addressed through ensuring that participants do not adopt cognitive strategies that would systematically influence experimental results (Leopold et al, 2001), to our knowledge, Experiment 3 is the first attempt to address decision criterion issue experimentally. In the domain of visual cognition, it has been difficult to disentangle whether effects are perceptual versus post-perceptual, particularly in light of theoretical controversies surrounding the extent to which visual perception and cognition are continuous (Pylyshyn, 1998). In categorical perception, a post-perceptual shift of category boundary can be detected as a shift in discrimination peak in a same-different task (Liberman, Harris, Hoffman & Griffith, 1957). This technique could not be employed for our global scene properties, however, as they are continuous dimensions. Signal detection theory has also been employed for determining changes in decision criterion (Swets, 1998), but these methods require that the signal be in a defined category, rather than values along a continuous dimension.

The strength of the adaptation paradigm is that it allows one to probe visual properties that are difference from, but that may depend on the adapted property. For example, Fang, Ijichi & He (2007) tested whether the coding of face viewpoint is independent of face identity and face gender using a transfer paradigm in which participants were adapted to an individual face at a particular viewpoint, and then asked to identify the viewpoint direction of a test face that could be either the same or different individual, or same or different gender as the adaptor face. This study found evidence of joint coding as adaptation did not completely transfer over gender or identity. The current experiment suggests that the perception of *openness* influences the rapid categorization of scenes as forests or fields, implying that basic-level categorizations might be mediated through the computation of structural properties such as *openness*. If this is the case, then we would expect that the categorization of prototypical forests and fields to also be modulated by the observers' adapted state to *openness*. We directly tested this hypothesis in Experiment 4.

Experiment 4: Adaptation to Openness Modulates Rapid Scene Categorization

Experiment 3 demonstrated that adaptation to a global property can change the classification of basic-level categories: exposure to closed or open scenes can change whether an ambiguous image would be classified as a member of the forest or field categories. This result suggests that openness may play a role in the rapid categorization of natural images as forests or fields. If the perception of global scene properties such as *openness* is necessary for rapid and accurate basic-level categorization then an observer's adapted state should change the speed and accuracy of prototypical scene categorization.

This was explored in Experiment 4. As in Experiment 3, participants in Experiment 4 were first adapted to streams of *open* and *closed* scenes. Following adaptation, they performed a basic-level categorization task on pictures of prototypical forests and fields. If the perception of *openness* is part of the scene representation allowing rapid basic-level categorization, then we predict the following cross-over interaction: participants should be slower and less accurate in categorizing fields after adapting to *open* images, and slower and less accurate in categorizing forests after adapting to *closed* images.

Methods

Participants

Ten participants (6 new, and 4 who had participated in Experiments 1, 2 or 3) participated in this experiment. All were between 18-35 years old and had normal or corrected-to-normal vision. Participants provided informed consent and were paid $10/h for their time.

Materials

The adaptation images in this experiment were the same images used in Experiment 3. The images used at test were 30 prototypical forests and 30 prototypical fields. The prototypicality of these scenes was determined from a previous ranking study (described in Experiment 3, with additional details in Greene & Oliva, 2009a). Images were determined to be prototypical if their mean ranking as forest or field was greater than 4 on a 5 point scale, and were not ranked as prototypical for any other scene category.

Procedure

Participants completed a two block experiment in which they were adapted to open and closed images in different blocks. Half of the participants adapted to *open* first, the other half to *closed* first. As we were only looking for an interaction in the experimental adaptation conditions, the control block of images was not used in this experiment. As in Experiments 1-3, each experimental block contained an adaptation phase and a test phase. The adaptation phase was identical to Experiment 3. In the test phase, participants performed a basic-level categorization task on prototypical forest and field images following the top-up RSVP adaptation before each trial. Participants were instructed to respond as quickly and accurately as possible as to whether the test image was a forest or a field. Because test images were prototypical exemplars of a scene category, visual response feedback was given (the word "Error" appeared on the screen for 300 ms following an incorrect categorization).

Results

For this experiment, we analyzed both reaction time and accuracy. Reaction times greater than two seconds were discarded from analysis. Data from one participant with mean RT of 2923 ms (group mean RT was 660 ms) was not included in the analysis. For the remaining participants, accuracy in this experiment was very high, approaching ceiling performance (accuracy average of 95%, median 96% correct). Therefore, the predicted interaction between scene category and adaptation condition was not observed ($F(1,8)<1$) for the accuracy data. However, for reaction times, we did observe a significant interaction between basic-level category and adaptation condition (F(1,8)=40.32, p<0.001). As shown in Figure 5B, observers were on average slower to categorize fields (average RT of 696 ms) than forests (average RT of 584 ms) after adapting to open images (t(8)=4.37, p<0.01). Adaptation to closed images did not have a significant effect on reaction time (average RT of 679 ms for fields, and 681 ms for forests).

Discussion

While Experiment 3 demonstrated that adapting to *open* or *closed* scenes could push the perception of novel ambiguous scenes towards being perceived as more field or forest-like, Experiment 4 went one step further, showing that the speed of categorization of prototypical forests and fields could be altered by the participants' adapted state to *openness*. For both Experiments 3 and 4, the effect is particularly strong for adaptation to *open*, rather than *closed* images. Together with the results of Experiment 3, the present results regarding a change in the speed with which prototypical images are categorized after adaptation, suggest a representational role for global properties in the rapid computation of a scene's basic-level category. As adaptation targets neural populations coding *openness*, the observed decrements in the speed of scene categorization can be taken as additional evidence of the openness property's role in representing these basic-level categories.

Importantly, Experiments 3 and 4 are the first behavioral evidence of a transfer of high-level semantic adaptation to an orthogonal task, providing critical insight into neural mechanisms that depend on the adapted property. In the case of natural image understanding, this provides a method for causally determining global scene properties that make up the representation of basic-level scene categories. Future work will involve elucidating which global scene properties participate in the representation of other basic-level scene categories (Greene & Oliva, 2009a).

General Discussion

Here we have demonstrated aftereffects to several global scene properties (Experiment 1). These aftereffects are not due to adaptation inherited from early visual areas (Experiment 2), and do not solely reflect a shift in the observers' decision criteria regarding the global scene properties (Experiment 3). Furthermore, we have demonstrated the perceptual consequences of global property adaptation to rapid scene categorization (Experiment 4), furthering the view that rapid scene analysis may be driven by the perception of such global properties (see also Greene & Oliva, 2009a).

Many of us have had the experience of traveling from our homes to a destination with very different visual features. For example, one might travel from a cold Boston winter to a sunny Florida beach. Upon returning from the trip, we might feel that our home is more gray and cold looking than remembered. Similarly, a city in the western United States might seem very open after visiting the dense and enclosed cities of the east coast. Such experiences demonstrate how our visual system adjusts to the input statistics of our current environment. In this laboratory demonstration we have shown that this process is rapid, and robust to changes in retinal position.

The use of adaptation and aftereffects has the potential to show important dimensions of stimulus coding. Webster et al (2004) demonstrated high-level aftereffects to the face dimensions of gender, expression and ethnicity. This served to be an important confirmation to the already accepted view that these dimensions were important to face coding. Scene understanding, on the other hand, does not yet have such readily accepted dimensions of coding. However, global properties such as *navigability* and *openness* have already been shown to be important dimensions of scene variability, as they are used by human observers in rapid scene categorization (Greene & Oliva, 2009a). They are therefore, reasonable properties for testing high-level aftereffects to environmental spaces. The existence of global property aftereffects, therefore gives considerable credence to a scene-centered view of scene recognition. Although we cannot fully reject the possibility that the neural axes of scene representation are not these global properties as we have defined them, but rather properties that are covariant with these properties, the presence of robust aftereffects to these global properties suggests that these dimensions of scene variability are important aspects of the semantic representation (or gist) of a scene. As adaptation directly targets populations of neurons coding a particular global property (Clifford 2005), the presence of aftereffects can be taken as evidence for the neural coding of such properties.

Although a variety of high-level aftereffects have been reported for faces (Leopold, O'Toole, Vetter & Blanz, 2001; Rhodes et al, 2005; Webster, 2004), relatively little work has been done investigating perceptual aftereffects to real-world scenes. One exception has been from Kaping and colleagues (2007). In this study, participants were adapted to texture patterns that had orientation distributions that were similar to either natural or urban images. Following adaptation, participants categorized moderately urban images as either natural or urban. They found that when the orientation statistics of the adapting textures matched natural scenes, the test images were more consistently classified as urban, and the reverse also being true for adapting images matching urban scene statistics. Our results are completely congruent with this study as we also found robust adaptation to naturalness using our paradigm. However, while the Kaping et al (2007) study demonstrates that adapting to a single image statistic alters the perception of scenes, our study demonstrates that considerable exposure to scenes with a specific set of global property regularities can alter the perception of subsequent scene images.

While adaptation to global scene properties had a significant effect for all measured properties in Experiment 1, the effect was unidirectional for *navigability*. As suggested by Figure 3E, adaptation to non-navigable environments seems to have an effect only on the least-navigable test images (25th ranking percentile). This leads to the possibility that it is not *navigability* that adapts per se, but rather information that is correlated with non-navigable environments. For example, very low navigability environments tend to be *closed* environments made up of dense textures (from elements such as thick brush or rock outcroppings), suggesting that the unilateral aftereffect could reflect adaptation to *closedness* or texture density (Durgin & Huk, 1997).

Experiment 2 demonstrated that the openness aftereffect cannot be explained by adaptation from early visual areas as the aftereffect was tolerant to a relatively large spatial shift across the vertical meridian. It is an open question of where in the visual system this adaptation takes place. However, a few general points can be made. While the eccentricity of our stimuli from the central fixation point is similar to the receptive field sizes reported to macaque V4 (Gattass, Sousa & Gross, 1988), our stimuli were presented on opposite sides of the vertical meridian and only IT has receptive fields that represent both hemifields (Gross, Rocha-Miranda & Bender, 1972), though the human homolog to this area is still an area of active research (Bell , Hadj-Bouziane, Frihauf, Tootell & Ungerleider, 2008).

The set of global properties used here was designed to describe major dimensions of natural scene variation, not to be an independent basis for describing scenes. There is some significant covariation existing between properties (Greene & Oliva, 2009a). In our experiments, attempts were made to test the properties as independently as possible. Our adaptation paradigm used a large number of real-world scenes that were selected to vary as much as possible in all spatial, semantic and low-level properties as possible while maintaining a consistent rank along the particular global property dimension.

In the domain of face processing, the concept of a "face space" has been in the literature for some time (Turk & Pentland, 1991). This framework has been particularly influential because rather than encoding the local features of a face, such as eyes, nose and mouth, it represents global patterns of individual variation. This framework has allowed work to be done on high-level adaptation for faces by providing a continuous, high-dimensional space. A global scene property framework provides much of the same function: it describes large patterns of global variation over natural environmental categories in a continuous way, without the need to represent the individual objects that a scene contains. Adaptation provides a method for testing the psychological reality of candidate dimensions for this scene space. As Experiments 3 and 4 also demonstrated, adaptation provides a method for testing the utility of these candidate properties for scene tasks, such as basic-level category recognition.

References

Afraz, S., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. Vision Research, 48(1), 42-54.

Appelton, J. (1975). *The Experience of Landscape*. London: Wiley.

Attneave F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61,183–93.

Baddeley, R. (1997). The correlational structure of natural images and the calibration of spatial representations. Cognitive Science, 21(3), 351-371.

Barlow HB. (1961). Possible principles underlying the transformation of sensory Messages.In *Sensory Communication*, ed. WA Rosenblith, pp. 217–34. Cambridge, MA: MIT Press.

Bell, A. H., Hadj-Bouziane, F., Frihauf, J. B., Tootell, R. B. H., & Ungerleider, L. G. (2009). Object Representations in the Temporal Cortex of Monkeys and Humans as Revealed by Functional Magnetic Resonance Imaging. Journal of Neurophysiology, 101(2), 688-700.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 443-446.

Bullier, J. (2004). Communications between cortical areas of the visual system. In L.M. Chalupa & J.S. Werner (Eds.), *The Visual Neurosciences* (pp. 522–540). Cambridge, MA: MIT Press.

Clifford, C. W., Wenderoth, P., & Spehar, B. (2000). A functional angle on some after-effects in cortical vision. *Proceedings. Biological Sciences / The Royal Society, 267*(1454), 1705-10.

Clifford, C. W. G. (2005). Functional ideas about adaptation applied to spatial and motion vision. In C. W. G. Clifford & G. Rhodes (Eds.), *Fitting the mind to the world: Adaptation and after-effects in high-level vision* (pp. 47–82). Oxford, UK: Oxford University Press.

Clifford, C. W. G., Webster, M., Stanley, G., Stocker, A., Kohn, A., Sharpee, T., et al. (2007). Visual adaptation: neural, psychological and computational aspects. *Vision Research*, *47*, 3125-3131.

Fang, F., Ijichi, K., & He, S. (2007). Transfer of the face viewpoint aftereffect from adaptation to different and inverted faces. *Journal of Vision*, 7(13), 1-9.

Fei-Fei, L., & Perona, P. (2005). A Bayesian Hierarchical model for learning natural scene categories. *IEEE Proceedings in Computer Vision and Pattern Recognition*, 2, 524-531.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), 1-29.

Gattass, R., Sousa, A., & Gross, C. (1988). Visuotopic organization and extent of V3 and V4 of the Macaque. *Journal of Neuroscience*, 8(6), 1831-1845.

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

Greene, M.R., & Oliva, A. (2009a). Recognition of natural scenes from global Properties: Seeing the forest without representing the trees. *Cognitive Psychology* 58(2), 137-176.

Greene, M.R., & Oliva, A. (2009b) The Briefest of Glances: the Time Course of Natural Scene Understanding. *Psychological Science* 20(4), 464-472.

Gross, C., Rocha-Miranda, C., & Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 35, 96-111.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 604-610.

Ito M, Tamura H, Fujita I, Tanaka K (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology, 73*:218-226.

Jiang, F., Blanz, V., & O'Toole, A. (2006). Probing the visual representation of faces with adaptation: a view from the other side of the mean. *Psychological Science*, 17(6), 493-500.

Joubert, O., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. *Vision Research*, 47: 3286-3297.

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition*, 15(1), 12-19.

Kaplan, S. (1992). Environmental Preference in a Knowledge-Seeking, Knowledge-Using Organism. In J. H. Barkow, L. Cosmides, and J. Tooby (Eds.) *The Adaptive Mind.* New York: Oxford University Press, 535-552.

Konkle, T., Wang, Q., Hayward, V., & Moore, C. I. (in press). Motion Aftereffects Transfer Between Touch and Vision. *Current Biology.*

Large, M.E., Culham, J., Kuchinad, A., Aldcroft, A., & Vilis, T. (2008). fMRI reveals greater within- than between-hemifield integration in the human lateral occipital cortex. *European Journal of Neuroscience,* 27(12), 3299-3309.

Leopold, D., O'Toole, A., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, 4, 89-94.

Liberman, A., Harris, K., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.

Loftus, G. R. & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, 1, 476-490.

Melcher, D. (2005). Spatiotopic transfer of visual-form adaptation across saccadic eye movements. *Current Biology*, 15, 1745-1748.

Motoyoshi, I., Nishida, S., Sharan, L, & Adelson, E. (2007). Image statistics and the perception of surface qualities. *Nature*, 447, 206-209.

Oliva, A., & Schyns, P. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176-210.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42, 145-175.

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

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

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavior and Brain Sciences*, 22, 341–423.

Rhodes, G., Robbins, R., Jaquet, E., McKone, E., Jeffery, L., & Clifford, C. W. G. (2005). Adaptation and face perception: How aftereffects implicate norm-based coding of faces. In C. W. G. Clifford & G. Rhodes (Eds.), *Fitting the mind to the world: Adaptation and after-effects in high-level vision* (pp. 213–240). Oxford, UK: Oxford University Press.

Rhodes, G., Jeffery, L., Clifford, C. W. G., & Leopold, D. A. (2007). The timecourse of higher-level face aftereffects. *Vision Research*, 47(17), 2291-6.

Rousselet, G. A. Joubert, O. R. Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, 12(6), 852-877.

Rousselet,G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision*, 3(6):5, 440-455

Swets, J.A. (1998). Separating discrimination and decision in detection, recognition and matters of life and death. In D. Scarborough & S. Sterngerg (Eds) *An Invitation to Cognitive Science: Methods, Models and Conceptual Issues* (p.635-702). Cambridge: MIT Press.

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Pattern Analysis and. Machine Intelligence*, 24, 1226-1238.

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

Troje, N. F., Sadr, J., Geyer, H., & Nakayama, K. (2006). Adaptation aftereffects in the perception of gender from biological motion. *Journal of Vision*, *6*(8), 850-857.

Vogel, J., & Schiele, B. (2007). Semantic scene modeling and retrieval for content-based image retrieval. *International Journal of Computer Vision*, 72(2), 133-157.

Wade, N. & Verstraten, F. (2005). Accommodating the past: a selective history of adaptation. In C. Clifford & G. Rhodes (Eds) *Fitting the Mind to the World: Adaptation and after-effects in high-level vision.* (p. 83-102) New York: Oxford University Press.

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

Webster, M. A. (1996). Human colour perception and its adaptation. *Network: Computation in Neural Systems, 7,* 587–634.

Webster, M., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural face categories. *Nature*, 428(6982), 557-561.

Winawer J, Huk A, Boroditsky L. (2008) A motion aftereffect from viewing still photographs depicting motion. *Psychological Science*. 19(3): 276-283.

Xu, H., Dayan, P., Lipkin, R. M., & Qian, N. (2008). Adaptation across the Cortical Hierarchy: Low-Level Curve Adaptation Affects High-Level Facial-Expression Judgments. *J. Neurosci.*, *28*(13), 3374-3383.

Chapter 5: Conclusion

In this thesis, I have shown a variety of behavioral evidence for a global scene-centered approach to natural scene understanding. This approach uses a small vocabulary of global and ecologically relevant scene primitives that describe the structural, constancy and functional aspects of scene surfaces without representing objects and parts; asserting that one may recognize the "forest" without necessarily first representing the trees.

In Chapter 2, human performance at a rapid scene categorization task was dramatically influenced by varying the distractor set to contain more global property similarities to a target category, suggesting that human observers were sensitive to global property information when performing rapid scene categorization (Chapter 2, Experiment 2). To what extent is global property information alone a sufficient predictor of rapid natural scene categorization? The performance of a simple classifier representing only these properties is indistinguishable from human performance in a rapid scene categorization task in terms of both accuracy and false alarms (Chapter 2, Experiment 3). To what extent is this high predictability unique to a global property representation? I compared two models that represented scene object information to human categorization performance and found that these models had lower fidelity at representing the patterns of performance than the global property model (Chapter 2, Experiment 4).

The time course of global property and basic-level category perception was explored in Chapter 3. If the initial scene representation contains substantial global property information that allows basic-level categorization, then observers should require less image exposure to correctly classify a scene's global property than to categorize it at the basic level. Indeed, I found that observers needed to see an image for less time on average to classify the scene's global properties than to categorize it at the basic level.

This result suggests the intriguing possibility that there exists a time during early visual processing where a scene may be classified as a large space or navigable, but not yet as a mountain or lake. This work is unique in that it compares the perceptual availability of a relatively large number of classification tasks (14). Comparing the relative availability of these different tasks can reveal bottlenecks in the accumulation of meaning. Understanding these bottlenecks provides critical insight into the computations underlying rapid visual understanding. Furthermore, given the extraordinarily rapid nature of some classifications (75% thresholds as little as 19 ms for *naturalness*), this result provides strong time constraints for early visual mechanisms of scene perception.

Last, Chapter 4 used an adaptation paradigm to explore the susceptibility of global properties to aftereffects and used the presence of aftereffects as a method to probe for a causal link between global property perception and the perception of the scene's basic-level category. In this chapter, I demonstrated aftereffects to several global scene properties (Chapter 4, Experiment 1). This work is the first laboratory demonstration of aftereffects from prolonged viewing of natural scene images. These aftereffects are not due to adaptation inherited from early visual areas (Chapter 4, Experiment 2), and do not solely reflect a shift in the observers' decision criteria regarding the global scene properties (Chapter 4, Experiment 3). This experiment provides a possible control experiment method for other work on high-level aftereffects as the potential for criterion shift is sometimes addressed in the literature as a potential weakness of high-level adaptation paradigms, but has not been experimentally addressed. I lastly demonstrated the perceptual consequences of global property adaptation to rapid scene categorization (Chapter 4, Experiment 4), showing systematic reaction time differences as a function of adaptation. As adaptation targets neural populations coding *openness*, the observed decrements in the speed of scene categorization can be taken as additional evidence of the openness property's role in representing these basic-level categories.

Taken together, the experimental results described in this thesis provide converging behavioral evidence for an initial global scene-centered visual representation. However, there are some limitations to the current approach that must be addressed. First, although the global properties used in this work have been found to influence human observers' basic-level scene categorization, I cannot make the claim that these global properties are exactly the ones being processed by the brain to allow categorization to take place. Rather, there is still the possibility that the brain is processing properties that are covariant with the currently defined properties. Similarly, as robust correlations exist between certain pairs of global properties, it could be that neural axes reflect aspects of more than one global property. One solution to this issue is to use functional brain imaging to determine the relevant scene axes, using either adaptation or pattern classification techniques, as both allow inferences to be made about the neural similarity of different stimuli.

A second limitation is that I cannot currently predict a scene's global properties from the image pixels, but rather only through the rankings of human observers. Finding the image features responsible for these global properties would be a great leap forward in this work. While image statistic correlates exist for some of the spatial global properties (Oliva & Torralba, 2001; Torralba & Oliva, 2002), I believe that a more fundamental question is which image features are human observers using to classify a scene's global properties? Hopefully, future advances would allow the use of reverse correlation techniques to be used to probe this question (Ahumada, 2002; Gosselin & Schyns, 2001).

As all global properties used here made some contribution to basic-level scene categorization, one must ask whether there is anything "special" about these properties or whether any reasonable description of a scene will do. Although Chapter 2 Experiment 4 shows that a scene description using objects fails to replicate human categorization errors, we are still left with the question of whether any global property would contribute in the computation of a scene's basic-level category. One possibility is that global properties that are "accidental" (in other words, not distinguishing between basic-level categories) will not contribute to scene categorization. For example, *clutter* is a dimension where many scene categories can vary – there can be more or less cluttered bedrooms and offices, for example. Another example might be mirror *symmetry* as this is a property that could depend more on the angle of the photograph than any intrinsic geometry of the space. Symmetry has been found to not be used by human participants in scene recognition (Sanocki & Reynolds, 2000). Psychophysical aftereffects provide a possible method for testing potential useful aftereffect as adaptation directly targets populations of neurons coding a particular global property (Clifford 2005). Therefore, presence of aftereffects to a candidate property can be taken as evidence for the neural coding of that property.

The global properties presented here work for natural environments, which reflect only a small subset of the scenes that we experience in our lives. A clear future extension of this approach would be to test the role of global properties for other types of environments, such as indoor scenes. While a corridor, for example, has a stereotyped spatial layout including a great deal of perspective, some of the current global properties used for natural environments may not be diagnostic of indoor environments. For example, no indoor environments are *open* and all are designed to permit *navigation*. This leaves us two possibilities for indoor scene recognition: (1) indoor scenes can be described using a different set of global properties; or (2) indoor scenes are recognized primarily through one or more prominent objects. The first possibility is testable, although new global properties specific to indoor scenes must be devised. For example, the maximum *occupancy* of the place is an intuitive functional property of indoor environments. A closet or bathroom would have smaller occupancy than a conference room or classroom. Although occupancy would increase with increased volume of the room, a bedroom of similar volume will have a smaller occupancy than a living room. Another global scene property for indoor scenes could describe the location of the scene's center of mass: a dining room or conference room has a more central mass than a kitchen or corridor. However, indoor scenes could also be well-described by an object model similar to those described in Chapter 2 Experiment 4. An empty room in a new house becomes a bedroom, an office, a library or a music studio depending on the objects that are placed in the room. Intuitively, the prominent object model seems like it would achieve high categorization performance on some indoor categories such as bedrooms or living rooms because the largest object (bed or sofa) is not typically found in other scene categories. Future work should examine the representations building all types of scene categories.

One last question surrounds the time course of object processing in building scene identity. Surely, as objects can make up the identity of the scene and are the entities acted on by agents in a scene, they are of critical importance for scene understanding with longer image exposures, but when and how does object information become available? This and other work emphasize that objects (especially small ones) might not be available for report at the beginning of the glance (Fei-Fei et al, 2007; Gordon, 2004; Rayner Smith, Malcolm & Henderson, 2008). Therefore, a critical question for scene understanding involves examining how object identity becomes available in the scene representation. As objects can vary in size and salience, answers about the availability of "objects" in general may be impossible. To make matters worse, it will be difficult to disentangle the perception of an object with inference (overestimation of what was seen due to the covariance with other perceived features and objects, see Brewer & Treyans, 1981). Therefore, a complete understanding of the representation of objects in scenes will require knowledge about object size and context. However, with large databases and object labeling techniques available on the internet, such as LabelMe (Russell, Torralba, Murphy & Freeman, 2008), it is now possible to gather these statistics and design the experiments. A view of the time course of object understanding within a scene, combined with the current work would provide a rich picture of the early dynamics of the human visual system.

Concluding remarks

All together, the results in this thesis provide support for an initial scene-centered visual representation built on conjunctions of global properties that explicitly represent scene function and spatial layout, but not necessarily the objects in the scene. This fills a critical gap in the literature on high level visual processing, by allowing a global scene representation to be operationalized and tested. It also presents a significant departure from traditional behavioral and modeling work on scene understanding which builds the scene from pixels to contours through objects and then finally the scene. Here, scene recognition can proceed without the laborious segmentation and object recognition stages, providing a novel account of how human observers could identify a place as a "forest", without first having to recognize the "trees".

References

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision, 2*(1), 121–131.

Brewer, W., & Treyans, J. (1981). Role of schemata in memory for places. *Cognitive Psychology*, 13(2), 207-230.

Clifford, C. W. G. (2005). Functional ideas about adaptation applied to spatial and motion vision. In C. W. G. Clifford & G. Rhodes (Eds.), *Fitting the mind to the world: Adaptation and after-effects in high-level vision* (pp. 47–82). Oxford, UK: Oxford University Press.

Fei Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), 1-29.

Gordon, R. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 760-777.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition. *Vision Research, 41,* 2261–2271.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42, 145-175.

Rayner, Smith, Malcolm & Henderson. (2008) Eye Movements and Visual Encoding During Scene Perception. *Psychological Science*, 20(1): 6-10.

Russell, B.C., Torralba, A., Murphy, K.P., & Freeman, W.T. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3), 157-173.

Sanocki, T. and Reynolds, S. 2000. Does figural goodness influence the processing and representation of spatial layout. *Investigative Ophthalmology and Visual Science*, 41:723.

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Pattern Analysis and. Machine Intelligence*, 24, 1226-1238.