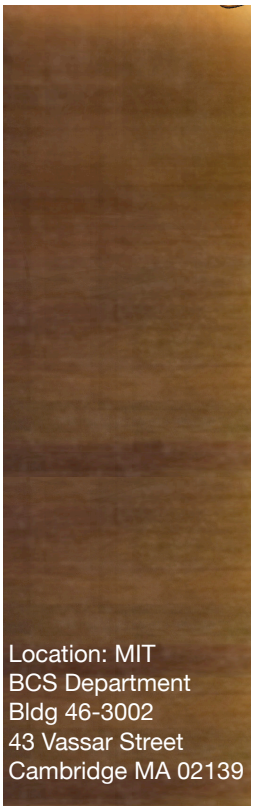
January 30, 2009



Location: MIT
BCS Department
Bldg 46-3002
43 Vassar Street
Cambridge MA 02139


The Scene Understanding Symposium series is an educational initiative from the Department of Brain and Cognitive Sciences, the McGovern Institute for Brain Research and the Computer Science and Artificial Intelligence Laboratory at MIT. We wish to thank for their sponsoring: the McGovern Institute for Brain Research, the National Science Foundation (IIS-CAREER Award to Aude Oliva, No. 0546262, and IIS-CAREER Award to Antonio Torralba, No. 0747120), the National Institutes of Health Graduate Training Grants (T32 EY013935 and T32 GM007484), the Department of Brain and Cognitive Sciences and the Computer Science and Artificial Intelligence Laboratory at MIT.

# Poster Session

1 **Learning to Describe Objects**

Ali Farhadi, Ian Endres, Derek Hoiem, David Forsyth
Computer Science, University of Illinois at Urbana-Champaign (UIUC)

We propose to shift the goal of recognition from naming to describing. Doing so allows us not only to name familiar objects, but also: to report unusual aspects of a familiar object ("spotty dog", not just "dog"); to say something about unfamiliar objects ("hairy and four-legged", not just "unknown"); and to learn how to recognize new objects with few or no visual examples. Rather than focusing on identity assignment, we make inferring attributes the core problem of recognition. These attributes can be semantic ("spotty") or discriminative ("dogs have it but sheep do not"). Learning attributes presents a major new challenge: generalization across object categories, not just across instances within a category. In this paper, our contributions include 1) methods for learning attributes that generalize well across categories; 2) methods for attribute-based learning from few examples; 3) methods for reporting unusual attributes of objects; 4) methods for learning new categories from textual description and 5) thorough evaluation that provides insights into the limitations of the standard recognition paradigm of naming and demonstrates the new abilities provided by our attribute-based framework.

## 2 Does gravity matter? Effects Of Semantic And Syntactic Object-Scene Inconsistencies On The Allocation Of Attention During Naturalistic Scene Viewing

Melissa Le-Hoa Võ & John M. Henderson
Psychology Department, University of Edinburgh, U.K.

There is evidence that attention and eye movements during scene viewing are preferentially allocated to semantically inconsistent objects compared to their consistent controls. However, there has been a dispute over how early during scene viewing such inconsistencies affect eye movement control. While the classical "octopus in a farmyard" study by Loftus and Mackworth (1978) and more recent studies by Underwood and colleagues (e.g., Underwood, Humphreys, & Cross, 2007) argue for extrafoveal detection of object-scene inconsistencies, other work has failed to find such early effects of semantic inconsistency on eye movement control (e.g., Henderson, Williams, & Hollingworth, 1999; Gareze & Findlay, 2007). The study presented here extends previous work by using highly controlled 3D-rendered images of real-world scenes instead of line drawings or photographs, which are either less realistic or more difficult to control for bottom-up saliency. In addition, we directly compared the effects of both semantic and syntactic object-scene inconsistencies on eye movement control during scene viewing. We therefore introduced syntactic object-scene inconsistencies (i.e., floating objects) in addition to semantic inconsistencies to investigate the degree to which they attract attention during scene viewing. In Experiment 1 participants viewed scenes in preparation for a subsequent memory task, while in Experiment 2 participants were instructed to search for target objects. In neither experiment were we able to find evidence for extrafoveal detection of either type of inconsistency. However, upon fixation both semantically and syntactically inconsistent objects led to increased object processing as seen in elevated gaze durations and number of fixations. Interestingly, the semantic inconsistency effect was diminished for floating objects, which suggests an interaction of semantic and syntactic scene processing. This study is the first to provide evidence for the influence of syntactic in addition to semantic object-scene inconsistencies on eye movement behavior during real-world scene viewing.

## 3 Capturing scene regularities with probabilistic geometric grammars

Meg Aycinena Lippow, Leslie Pack Kaelbling, Tomas Lozano-Perez
CSAIL, MIT

Images of similar scenes usually exhibit strong patterns in the types and spatial arrangements of objects, as many researchers have noticed. However, very few object detectors incorporate knowledge of the expected relationships among objects into the detection process. This may be due to the complexity of these relationships; there may be one or several desks in an office; a kitchen may contain a table, or it may not; and the fork could be to the left or to the right of the plate in a placesetting. We propose to capture these complex scene regularities using probabilistic geometric grammars (PGGs). A simple but flexible class of models, PGGs can model hierarchical groupings of objects within and across similar scenes, and capture variability in the number and combinations of objects through distributions over rules. Furthermore, they can naturally represent conditional independences among subgroups with the context-free assumption. We have developed an approach for learning the structure and parameters of PGGs for scenes from data, and a simple and efficient detection algorithm given a learned grammar. We show preliminary detection results, and compare them to results using object detectors alone.

## 4 Developmental Changes of the Encoding of Indoor versus Outdoor scenes

Xioaqian J. Chai, Noa Ofen, Lucia F. Jacobs, John D.E. Gabrieli
Brain and Cognitive Sciences, MIT; Department of Psychology, UC Berkeley

Regions in the medial temporal lobe (MTL) including the parahippocampal cortex (PHC) are involved in memory formation. Previously we showed that activations in MTL for successful memory encoding of indoor and outdoor scenes does not change with age (Ofen et al., 2007). Indoor and outdoor scenes, however, differ in their statistical properties and complexity. Moreover, in adults, viewing of indoor scenes is associated with greater activation in the posterior PHC when compared to viewing outdoor scenes (Henderson et al., 2007). The development of brain activation for processing of indoor versus outdoor scenes is unknown although; there is evidence of developmental changes in brain regions involved in the processing of scenes (Golarai et al., 2007). Furthermore, the potential contribution of these activations to memory formation is unclear. Here we assessed the development of brain regions for processing of indoor and outdoor scenes in the context of subsequent memory formation in 49 participants (ages 8 - 24). We found that memory for indoor, but not outdoor, scenes increased with age. Overall, indoor scenes compared to outdoor scenes activated bilateral posterior PHC and retrospelenial cortex and, within these regions, the activation in the right posterior PHC increased with age. Using a complexity index we further found that activation in the right posterior hippocampus increased with age for more complex indoor scenes. Collectively, these findings suggest prolonged developmental trajectory of posterior MTL regions and their roles in successful memory recollection.

## 5 Semantic guidance of eye movements during real-world scene perception

Alex Hwang, Hsueh-Cheng Wang, Marc Pomplun
Department of Computer Science, University of Massachusetts at Boston

Real-world scenes are filled with real-world objects that have not only visual representations, but also have meanings and semantic relations among them. The guidance of eye movements based on visual appearance (low-level visual features) has been well studied in terms of both bottom-up and top-down aspects. However, effects on eye movements by object meaning and object relations have been studied comparably less because of a few hurdles that make such study more complicated: (1) Object segmentation is difficult, (2) Semantic relations among objects are hard to define, and (3) A quantitative measure of semantic guidance has to be developed. In this study, semantic guidance is measured for the first time, thanks to the efforts made by two other research groups: first, the development of the LabelMe annotated image database (Russell et al., 2008), and second, the text/word latent semantic analysis tool which computes the cosine of the angle between the vectors representing two terms in semantic space, the LSA@CU (Landauer et al., 1998). We were able to generate a series of semantic salience maps following each eye fixation which approximates the transition probability to the next object assuming that eye movements are entirely guided by the semantic relations between objects. Subsequently, these semantic salience maps are combined and the final strength of guidance is measured by the Receiver Operator Characteristic (ROC), which computes the extent to which the actual eye fixations followed the ideal semantic salience map. Our analysis confirms that eye movements during visual inspection are strongly guided by a semantic factor based on the conceptual relations between the currently fixated object and the target object of the following saccade, indicating semantic guidance. The function of this guidance may be to ensure the consecutive inspection of conceptually closely related objects in order to facilitate memorization of the scene for later recall.

## 6 A model of top-down attentional control during visual search in complex scenes

Alex Hwang, Emily C. Higgins &, Marc Pomplun
Department of Computer Science, University of Massachusetts at Boston

Recently, there has been great interest among vision researchers in developing computational models that predict the distribution of saccadic endpoints in naturalistic scenes. In many of these studies, subjects are instructed to view scenes without any particular task in mind so that stimulus-driven (bottom-up) processes guide visual
attention. However, whenever there is a search task, goal-driven (top-down) processes tend to dominate guidance, as indicated by attention being systematically biased towards image features that resemble those of the search target. In the present study, we devise a
top-down model of visual attention during search in complex scenes based on similarity between the target and regions of the search scene. Similarity is defined for several feature dimensions such as orientation or spatial frequency using a histogram-matching technique. The amount of attentional guidance across visual feature dimensions is predicted by a previously introduced informativeness measure. We use eye-movement data gathered from participants' search of a set of naturalistic scenes to evaluate the model. The model is found to predict the distribution of saccadic endpoints in search displays nearly as accurately as do other observers' eye-movement data in the same displays.

## 7 The VideoIQ iCVR: Embedding scene analysis and intelligent video storage in the camera

Dimitri Lisin, Aleksey Lipchin, Igor Reyzin, and Mahesh Saptharishi
VideoIQ, Inc.

We present the VideoIQ iCVR: a smart camera with built-in video recording and analytics. The iCVR is able to reliably detect and track objects of interest and generate alerts when user-defined conditions ("rules") are violated. Results of the scene analysis intelligently control the quality and amount of video that is stored on the camera. Compression quality levels are adjusted based on the content of the video stream. Stored video can be browsed quickly based on content rather than linear playback. The system works in real time, is robust in highly dynamic environments, and is capable of handling different video signal types (e. g. color, grayscale, thermal IR and near IR). Unlike typical background subtraction techniques, the VideoIQ approach to scene analysis is characterized by its extensive use of feedback from high-level (object classification) to low-level (focus-of-attention) processes. This feedback mechanism allows a unified approach that combines the computational simplicity of background modeling with the superior detection accuracy of a trainable cascade of object detectors. Moreover, feedback allows the system to be driven by object-level rather than pixel-level decisions. A signature of each detected and tracked object of interest is used for both appearance based tracking and forensic search. The signature is a rather simple, but surprisingly effective histogram in a modified hue-saturation-intensity color-space.

## 8 Natural scene categorization by global scene properties: Evidence from patterns of fMRI activity

Soojin Park, Michelle R. Greene, Timothy F. Brady, Aude Oliva
Department of Brain & Cognitive Sciences, MIT

Human scene categorization is remarkably rapid and accurate, but little is known about the neural representation mediating this feat. While previous studies on neural representation of scenes have focused on basic level scene categories, here we examined whether the neural representation of scenes reflect global properties of scene structure, such as openness of a space, or properties of surfaces and contents within a space, such as naturalness. In an fMRI study, human participants performed a one-back task on blocks of images of four scene groups: Open Natural images, Closed Natural images, Open Urban images, Closed Urban images. Each image group included multiple basic level categories. For example, Open Natural images included open views of fields, oceans and deserts; while Open Urban images included open views of highways, parking lots, and airports. For each participant, we defined regions of interest (ROIs) of the parahippocampal place area (PPA), the fusiform face area (FFA), lateral occipital complex (LOC) and V1. Multivariate pattern analysis was applied to voxels within each ROIs, and split-half pattern correlation and Euclidian distances across voxel activations were calculated (Haxby et al., 2001). We observed high identification accuracy in the PPA and V1, but not in the FFA and LOC. Most interestingly, when the correct identification failed in the PPA, the confusion was between images with the same layout rather than between images with the same content. For example, Open Natural images were often highly correlated with Open Urban images, but rarely with Closed Natural images. These results suggest that a critical component of scene representation in the brain is the coding of global properties of spatial layout.
Acknowledgement: Funded by NSF CAREER award to A.O. (IIS 0546262) and NSF-GRF to M.R.G. and T.F.B.

## 9 Investigating the relationship between human scene similarity perception and feature-space distances

Michael G. Ross, Emmanuelle Boloix, Aude Oliva
Department of Brain & Cognitive Sciences, MIT

Vision researchers have developed many visual features for use in computer classification algorithms and human perceptual models. Although these features have been compared and validated by measuring their utility with respect to particular visual tasks, their relationship to humans' generic perceptions of visual similarity are not well understood. To address this issue, we conducted an experiment in which target outdoor scene images were drawn from a large database and displayed alongside their ten nearest neighbors in six different feature spaces: two versions of Oliva & Torralba's Gabor scene gist, Lazebnik et al.'s SIFT grids, Tieu & Viola's filter cascade, a similar Gabor-filter cascade, and color spatial histograms. Human observers indicated which neighbors were semantically similar to the target. A first analysis of this data indicate that different features work best on different target classes. Additionally, on many targets multiple features are successful, but retrieve different images from the database. These developments make us optimistic that an approach that adaptively combines all feature types can lead to a superior algorithm for image similarity measurement and image database retrieval.
Acknowledgement: Funded by NSF grant to A.O. (IIS 0546262).

## 10 Modeling Search for People in 900 Scenes: A combined source model of eye guidance

Krista A. Ehinger * (1), Barbara Hidalgo-Sotelo * (1), Antonio Torralba (2) & Aude Oliva
(1) Department of Brain and Cognitive Sciences, (2) Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

How predictable are human eye movements during search in real world scenes? We recorded 14 observers' eye movements as they performed a search task (person detection) in 912 outdoor scenes. Observers were highly consistent in the regions fixated during search, even when the target was absent from the scene. These eye movements were used to evaluate computational models of search guidance from three sources: saliency, target features, and scene context. Each of these models independently outperformed a cross-image control in predicting human fixations. Models that combined sources of guidance ultimately predicted 94% of human agreement, with the scene context component providing the most explanatory power. None of the models, however, could reach the precision and fidelity of an attentional map defined by human fixations. This work puts forth a benchmark for computational models of search in real world scenes. Further improvements in modeling should capture mechanisms underlying the selectivity of observer's fixations during search. Funded by a Singleton Fellowship and a graduate training fellowship (T32 EY013935) to K.A.E.; National Science Foundation Graduate Research Fellowship to B.H.S; NSF CAREER award (0546262) and a NSF contract (0705677) to A.O., and NSF CAREER award to A.T (0747120). * These authors contributed equally.

## 11 A Vision-based Navigation Assistant

Olivier Koch, Seth Teller
CSAIL, MIT

We describe a vision-based guidance system meant to assist a human in navigating through a previously explored environment. The system incorporates several novel elements, including an uncalibrated, body-worn multi-camera rig, and an algorithm that learns the correlation between feature motion and egomotion without inertial sensing or intrinsic or extrinsic camera calibration. It offers coarse, rather than precise, guidance to the user, which we show is sufficient for effective navigation within an interesting class of environments. We demonstrate the system operating in several real-world exploration sessions, and evaluate its performance through qualitative and quantitative means.

## 12 A Bayesian model of top-down visual attention

Sharat Chikkerur, Thomas Serre, Cheston Tan and Tomaso Poggio
McGovern Institute, MIT

Understanding how human observers attend to objects in complex natural images is an important part of understanding how the visual cortex processes visual scenes. In past studies, several cues have been shown to influence the deployment of attention and eye movements. How the visual system combines these cues and what the underlying neural circuits are, remain however largely unknown. We present a Bayesian framework to model the contributions of bottom-up exogenous cues as well as top-down endogenous (feature-based and contextual) cues during complex visual search tasks. In addition, the Bayesian framework models the interaction between the parietal cortex and ventral stream during a search task and is consistent with recent physiological studies (Bichot et al. 05, Buschman and Miller 07). Here we use the model to explain human eye-movements in a complex visual search task. We show that both feature-based and contextual cues in isolation predict human eye-movements more accurately than bottom-up cues. Overall, the proposed model

(combining feature-based and contextual cues) achieves 94% of human performance on two different search tasks (pedestrians and cars). We also compare the model with human performance on a masked 'Animal vs. Non-animal' rapid recognition task (Serre et al 07). We show that the increase in performance observed for human observers when the delay between the stimulus and the mask (i.e., the Stimulus Onset Asynchrony, SOA) is increased can be explained by an increase in the contributions of top-down attentional signals.

## 13 Visualizing LabelMe

Antonio Torralba and Bryan Russell
CSAIL, MIT

Currently computers have difficultly with recognizing objects in images. While practical solutions exist for a few simple classes, such as human faces or cars, the more general problem of recognizing all the different classes of objects in the world (e.g. guitars, bottles, telephones) remains unsolved. Computer vision researchers are currently investigating methods that can recognize and localize thousands of different object categories in complex scenes. A key component of these algorithms is the data used to train the computer's model of each object. The goal of LabelMe is to provide an online annotation tool to build a large database of annotated images by collecting contributions from many people. In this poster, we show an interactive visualization of all the annotated images in LabelMe.

## 14 Modeling Perceptual Organization in Higher Dimensional Space

Ruth Rosenholtz, Nadja Schinkel-Bielefeld, Nathaniel R. Twarog & Martin Wattenberg
Department of Brain & Cognitive Sciences, MIT

An important facet of human vision is its ability to seemingly effortlessly perform “perceptual organization”; it transforms individual feature estimates into perception of coherent regions, structures, and objects. We perceive regions grouped by proximity and feature similarity, grouping of curves by good continuation, and grouping of regions of coherent texture. We discuss a simple model for a broad range of perceptual grouping phenomena. It takes as input an arbitrary image, and returns a structure describing the predicted visual organization of the image. We demonstrate that this model predicts visual percepts in classic perceptual grouping displays. We also demonstrate the performance of this model on several graphic and user interface designs, which are particularly well suited to testing such an algorithm due to their complexity and the degree to which their perceptual organizations have been recorded and studied.

## 15 Where in the World? Human and Computer Geolocation of Images

James Hays & Alexei A. Efros
School of Computer Science , Carnegie Mellon University

Previously, both human and computer vision researchers have mostly been interested in scene classification as it relates to rigid semantic categories (e.g. kitchen, bedroom, forest, city, etc...). It turns out that in this regime, simple gist or texture features can classify scenes almost as well as humans (Torralba & Oliva 2003, Renninger 2004), even with a small amount of training data. However, the success of computational methods might simply be due to the small number of scene

categories, and the ease with which these hand-designed categories can be separated by low-level features. Here we would like to investigate a much harder task. In this study we examine human performance at organizing scenes according to geographic location on the Earth rather than hand-defined semantic categories. Participants are shown novel images and asked to pick the location on a globe where the photograph was taken. This task is difficult -- many scenes are geographically ambiguous while others require high-level scene understanding and knowledge of cultural or architectural trends across the Earth. We compare and contrast human performance with a data-driven computational method using 6.5 million geolocated photographs. For a novel photograph, the algorithm finds the most similar scenes according to the scene gist descriptor, texton histogram, and other features. A voting scheme produces a geolocation estimate from the locations of matching scenes.

## 16 Task-driven Saliency Using Natural Statistics (SUN)

Matthew H. Tong, Christopher Kanan, Lingyun Zhang, & Garrison W. Cottrell
Department of Computer Science and Engineering, University of California, San Diego

One important task of the visual attention system is to focus attentional resources on important objects in a scene. Starting with a probabilistic definition of this goal, we derive a salience model that directs attention to locations likely to contain objects of interest. The Saliency Using Natural Statistics model (SUN) utilizes three kinds of statistical knowledge about the world in choosing which areas of a scene should be fixated: what features are rare, the visual appearance of particular objects of interest, and the locations in a scene likely to contain such objects. The components that emerge all have been proposed individually before. Novelty of features has been argued to attract attention (see Wolfe, 2001 for a discussion). SUN's appearance model is reminiscent of Guided Search (Wolfe, 1994) and Iconic Search (Rao et al., 1995) and location-based guidance has also been argued to play an important role (e.g. Turano, 2003). Unlike other models of saliency, SUN learns its statistics from natural image statistics in advance from a collection of images of natural scenes. Previous work with SUN defined the self information of features (their novelty) as a form of bottom-up, task-independent saliency and demonstrated its state-of-the-art ability to predict human fixations when free viewing images (Zhang et al., in press) and video. SUN's use of natural statistics learned through experience also explains search asymmetries that models driven solely by the statistics of the current image cannot. However, when viewing the world we do so with a purpose, and understanding the role of the current task is essential. Here we go beyond the previous work, implementing the remaining portions of SUN and applying the complete model to a real world task. Torralba and colleagues collected data from subjects counting people, paintings, and cups in indoor and outdoor scenes and assess the performance of their contextual guidance model on this data (2006). We compare SUN's performance with contextual guidance, finding that SUN's use of learned statistics matches or exceeds the performance provided by contextual guidance. This comparison is of particular interest as the two probabilistic models share many surface similarities, with contextual guidance being strongly influenced by the statistics of the current scene and SUN relying more heavily on previous experience.