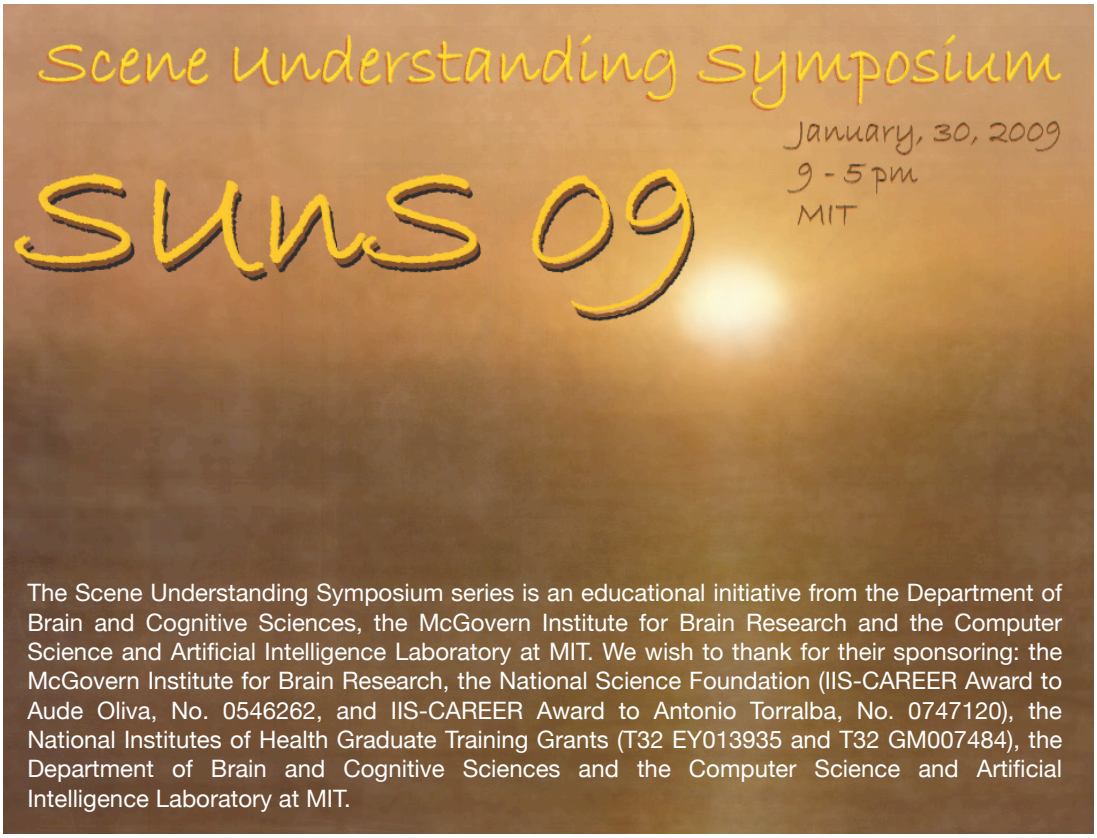
January 30, 2009



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.

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

Schedule of Talks

8:15 **Breakfast served**

8:40 **Opening remarks**

Session: Visual search and attentional guidance in scenes

8:45 **Context rules supreme in visual search through real-world scenes**

Aude Oliva, Krista Ehinger, Barbara Hidalgo-Sotelo & Antonio Torralba
Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology

How predictable are human eye movements as they search real world scenes? We recorded eye movements from observers searching through 1000 real world scenes for a pedestran and found a high degree of agreement in the regions fixated across observers, even when the scene lacked a target and the regions were not visually salient. In an effort to capture the guidance mechanisms driving this consistency, we modeled three sources of guidance (saliency, target features, and scene context) and evaluated how accurately they predicted human fixations. Each of the models independently outperformed a cross-image control, but it was the scene context module that provided the most explanatory power. In this large dataset, models that combined sources of guidance were able to predict more than 4 5 human search fixations. However, none of the models could reach the precision and fidelity of a human-based attentional map.

9:00 **A population code for selecting saccade targets during search**

Gregory Zelinsky
Psychology Department, SUNY Stony Brook

Theories of search often assume that overt attention moves directly from one object to the next, with the object of each fixation prioritized according to some measure of salience or similarity to a target. We present evidence for off-object fixations that challenges this view. Quite often eye movements land, not on objects, but rather near objects, between objects, or on background patterns en route to an object. We document this behavior using a diversity of stimuli, ranging from simple OQ displays to toys in a crib to fully realistic scenes, and interpret these fixations as evidence for population coding in the saccade selection process. To formalize this perspective a model is presented that uses a population code to describe these off-object fixations during search (Zelinsky, 2008). A search target and scene are coded in terms of simple linear filter responses, which are then correlated to create a map-based representation of target-scene similarity. Each point in this map votes (weighted by its activation) for a saccade to its location, with the centroid of this activation representing the population vote. The map is thresholded over time until the centroid achieves a criterion distance from the current fixation, at which time an eye movement is programmed to the centroid. This process continues until the target is detected or all non-target activity is pruned from the map, an event causing the simulated fovea to become aligned with the target. Simulated and human eye movements are compared for a variety of tasks and manipulations, and in all cases the agreement is quite high (typically within the 95% confidence interval). This suggests that a fixed-parameter model using spatio-chromatic filters and a population code can be a good predictor of human overt search behavior.

9:25 **SUN: A model of visual salience using natural statistics**

Garrison Cottrell
Department of Computer Science and Engineering, University of California San Diego

As a result of having a foveated retina, we actively move our eyes in order to direct our highest resolution of visual processing towards interesting things. In fact, we move our eyes about three times a second; it is a decision we make about 172,000 times a day. How do we decide where to look? Most computational theories of overt visual attention assume there is a salience map that is computed from both exogenous and endogenous sources. Exogenous sources of salience tend to be visually "busy" locations in the world. Endogenous sources of salience derive from task-relevant considerations, such as those involved in my most vexing task of the day, finding my glasses. We have developed a Bayesian model of salience that naturally leads to components corresponding to these two influences. Our model differs from most other models in assuming that for exogenous influences on salience, the statistics of the visual world are learned through experience and then applied to new stimuli, as opposed to being computed directly from new stimuli. The model has been applied to standard datasets and can account for eye fixations as well or better than any other model in the literature, is able to be computed in real time due to its computational efficiency, and is able to explain several visual search asymmetries.

9:50 **Coffee break**

10:10 **Bottom up and top-down guidance of visual attention in natural environments**

Lauren Itti
Computer Science Department, University of South California

The natural world affords a complexity which makes its comprehension highly complex. To cope with this complexity, biological systems have evolved attentional strategies which rapidly focus processing resources on the most important and relevant aspects of the incoming sensory data. Here I will describe several exciting new research directions that study the joint stimulus-driven (or bottom-up) and goal-driven (or top-down) influences on attentional allocation. I will describe a new computational model which processes video inputs and predicts where observers look under different task conditions. I will discuss results of testing this model against human eye movement recordings over several hours of video stimuli.

10:30 **Search in real scenes: The latest mysteries, the latest clues**

Jeremy Wolfe
Brigham & Women's Hospital, Harvard Medical School

In scenes where every pixel has been assigned to an object or named region, the number of regions can be used as an estimate of the set size. If we perform visual search for arbitrary objects in those scenes, reaction time (RT) is almost unrelated to the set size, defined in this manner. That was last year's SUNS talk. This year, we concede that, counting objects is simply the wrong approach to understanding the efficiency of search in scenes. It is instructive to compare search for objects in scenes and in small arrays on blank backgrounds. We had Os search for distinctive objects that were not in scenes. We used very small set sizes (1-4) and uncrowded displays. Nevertheless, search was very inefficient (about 55 msec/item). If we project the RTs from search in scenes onto the RT x set size function for simple objects, we see that search in our scenes is equivalent to search though 6-16 objects even though the scene might have 70 labelled "objects" in it (Call this the "effective set size"). Relatively small effective set size suggests that powerful guidance mechanisms limit the deployment of attention in scenes. The nature of that guidance is somewhat elusive. For example, the present data do not tell us if that guidance is object-based or scene-based.

10:50 **The influence of complex visual context, eye movements, and attention on visual processing and perception**

Michael Paradiso[1], Octavio Ruiz[1], Xin Huang[2], Sean MacEvoy[3], Cathy Clarke[4]
[1]Department of Neuroscience, Brown University; [2]Department of Physiology, University of Wisconsin;
[3]Center for Cognitive Neuroscience, University of Pennsylvania; [4]Beth Israel Deaconess Medical Center

Vision is fundamentally and profoundly contextual yet, for experimental rigor, many experiments study visual processing in the absence of visual and behavioral context. Our research demonstrates significant ways in which complex natural scenes, eye movements, and attention modify visual responses, receptive fields, and visual performance.

11:15 **Coffee break**

Session: Neural mechanisms of natural vision

11:30 **Physiological responses in human visual cortex during dynamic viewing conditions**

Gabriel Kreiman, Hesheng Liu, Yigal Agam and Joseph Madsen
Children's Hospital, Harvard Medical School

Natural scenes extend the typical battery of stimuli used in Vision (such as gratings and moving dots) in the space domain. Here we extend visual stimuli in the temporal domain. While most visual studies focus on static images, under natural viewing conditions the retinal image shifts constantly due to changes in the outside world and to head/eye movements. We record intracranial field potentials from subjects with intractable epilepsy. These recordings allow us to provide a millisecond-resolution view of the neural responses along the ventral visual stream while subjects view brief movie clips. We could accurately decode object category information in single trials as early as 100 ms post-stimulus. Decoding performance was robust to changes in rotation, scale and clutter. The results reveal that physiological activity in the human temporal lobe can account for some of the key properties of visual recognition and provide strong constraints for computational models of human vision.

11:50 **A hierarchy of temporal receptive windows in human cortex**

Uri Hasson
Department of Psychology and the Neuroscience Institute, Princeton University

Real-world events unfold at different time scales, and therefore cognitive and neuronal processes must likewise occur at different time scales. In the talk I will present a novel procedure that identifies brain regions responsive to the preceding sequence of events (past time) over different time scales. The fMRI activity was measured while observers viewed silent films presented forward, backward, or piecewise-scrambled in time. The results demonstrate that responses in different brain areas are affected by information that has been accumulated over different time scales, with a hierarchy of temporal receptive windows spanning from short (~4 s) to intermediate (~12 s) and long (~ 36 s). Thus, although we adopted an open-ended experimental protocol (free viewing of complex stimuli), we found that parametric manipulation of the temporal structure of a complex movie sequence produced lawful changes in cortical activity across different brain regions. We conclude that, similar to the known cortical hierarchy of spatial receptive fields, there is a hierarchy of progressively longer temporal receptive windows in the human brain.

12:15 **Poster spotlight**

12:40 **Lunch & posters**

Session: Visual routines for image understanding

2:00 **Correlated predictors, words and scenes**

David Forsyth
Univerity of Illinois at Urbana-Champaign

What kinds of scene are there? One kind is a setting in which particular kinds of object are likely to appear. We can't build a labelled data set, with some examples of each class of setting, because we don't know what the classes are. But pictures of the same scene should look similar to each other, and should contain comparable kinds of object. I will describe work that uses these ideas to build clusters of scenes from annotated image data. Another kind of scene is a geometric structure that helps determine what object hypotheses are plausible. I will describe work to infer such geometric contexts for complex indoor worlds. Yet a third kind represents the distribution of light in the world, to give us a cue to the shape and lightness of objects. I will point to this kind of scene as being poorly understood and worth understanding.

2:25 **Dense scene alignment**

Ce Liu, Jenny Yuen, and Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, (MIT)

We proposed SIFT flow that establishes dense, semantically meaningful correspondence between two images across scenes by matching pixel-wise SIFT features. Using SIFT flow, we develop a new framework for image parsing by transferring the metadata information, such as annotation and motion from the images in a large database to an unknown query image. In particular, we demonstrated a nonparametric scene parsing system using label transfer, with very promising experimental results suggesting that our system outperforms the state-of-the-art.

2:45 **Recursive compositional models for vision**

Alan Yuille
Department Statistics and Psychology, University of California - Los Angeles (UCLA)

Bayesian approaches to vision suggest that scene understanding can be formulated as image parsing where the goal is to decompose the image into its underlying patterns. One strategy is to perform inference by data driven Markov Chain Monte Carlo (Tu and SC Zhu 2002, Tu, Chen, Yuille, SC Zhu 2005) which is a bottom-up/top-down algorithm suggestive of the feedforward/feedback pathways in the brain. But to proceed further with this strategy requires models for visual patterns (e.g. objects, entire images) that can be represented compactly and which have efficient learning and inference algorithms. This motivates the design of recursive compositional models (RCMs) which has been successfully applied to a range of vision tasks (L Zhu. PhD Thesis 2008). RCMs build hierarchical models out of elementary components and have rapid inference algorithms. We have demonstrated different forms of learning ranging from fully unsupervised to supervised. The unsupervised learning automatically learns models for Gestalt-type image structures as well as models for objects. (Work with Long Zhu).

3:10 **How big is it?**

Daniel Kersten
Psychology Department, University of Minnesota

Whenever we navigate a scene, grasp an object, or even assess threat, size information is critical to appropriate actions. Yet the computational and neural solutions to size extraction are in large part unknown. There are two general problems. The first has been appreciated for centuries--a big far object and a small near object can project to the same size retinal image, so depth is a confounding variable. Further, extracting depth information from scenes is complex, potentially involving multiple sources of information from the context in which objects are seen. A second problem is determining the 2D size of the retinal image of an object. This requires finding an object's boundaries--a computational problem whose difficulty was not fully appreciated until the advent of computer vision. These two problems suggest that one might see evidence of informational coupling between lower-level cortical areas representing 2D retinotopic spatial information, such as V1, and higher-level regions associated with scene context and depth. I'll describe neuroimaging results showing that the 2D spatial extent of activity in primary visual cortex (V1) is modulated by depth information from a scene. More specifically, the spatial line response (fMRI BOLD) on V1, to a ring-shaped object, shifts anteriorly when the object appears bigger as a consequence of a change in scene context and thus apparent depth, even though the object's retinal size is unchanged.

3:35 **Looking at things and stuff**

Edward Adelson
Brain & Cognitive Sciences, Massachusetts Institute of Technology

Is material recognition a skill of its own, or is it derived from other recognition processes? Maybe it depends on object recognition: to recognize a wooden chair, we first recognize the chair and then the wood. Or maybe it depends on scene perception: to recognize a gravel road, we first recognize the road, and then the gravel. Using an image database of real world materials, we've run experiments indicating that material perception is a skill of its own. Among other things, it turns out that material recognition is really fast. You can categorize an image as plastic, paper, or fabric virtually as fast as you catergorize a stimulus as red or blue. Could it be that this rapid recognition just derives from low level image statistics (i.e., is material perception just 2D texture perception)? No, the images in our database have highly diverse appearances, which almost certainly cannot be captured with low level statistics. Material perception can, of course, be influenced by low level statistics, as well as by high level object or scene recognition. But it is a distinct skill that is highly developed in humans.

3:55 **Coffee break**

Session: Exploring virtual and dynamic environments

4:10 **The challenges of dynamic environments**

Mary Hayhoe
University of Texas at Austin

The sequential acquisition of visual information from scenes is a fundamental component of natural visually guided behavior. However, little is known about the control mechanisms responsible for the eye movement sequences that are executed in the service of such behavior. Theoretical attempts to explain gaze patterns have almost exclusively concerned two-dimensional displays that do not accurately reflect the demands of natural behavior in dynamic environments, or the importance of the observer's behavioral goals. A difficult problem for all models of gaze control, intrinsic to selective perceptual systems, is how to detect important, but unexpected stimuli without consuming excessive computational resources. We show, in a real walking environment, that human gaze patterns are remarkably sensitive to the probabilistic structure of the environment, suggesting that observers handle the uncertainty of the natural world by pro-actively allocating gaze on the basis of learnt statistical structure. This is consistent with the role of reward in the oculomotor neural circuitry, and supports a reinforcement learning approach to understanding gaze control in natural environments.

4:40 **The geometry of spatial knowledge for navigation**

William H. Warren
Dept. of Cognitive & Linguistic Sciences, Brown University

It is often assumed that "cognitive maps" in animals and humans have a Euclidean geometric structure. Such spatial knowledge might be built up from path integration, and would support accurate shortcuts and detours. However, apparently Euclidean behavior may also result from weaker knowledge together with adaptive navigation strategies. We use an ambulatory virtual environment to investigate active navigation in walking humans. Results indicate that humans have poor path integration, rely heavily on visual landmarks and topological (graph and neighborhood) structure, tolerate radical violations of Euclidean and ordinal structure, but can also fall back on coarse knowledge of distances and angles (weighted graph). Navigation appears to depend on strategies that exploit multiple forms of spatial knowledge, which are not integrated into a globally consistent map.

5:15 **Reception**