SUnS 06

Scene Understanding Symposium

Friday, February 17, 2006

Location	MIT Department of Brain and Cognitive Sciences Building 46 - 3002 (Auditorium)
8:40 am	**Coffee**
8:55	Opening Remarks
9:00-9:20	**From zero to gist in 200 msec: The time course of scene recognition** *Aude Oliva*, *Michelle Greene*, MIT Brain & Cognitive Sciences
9:20-9:45	**Feedforward theories of visual cortex predict human performance in rapid image categorization** *Thomas Serre & Tomaso Poggio*, MIT McGovern Institute
9:45-10:05	**Latency, duration and codes for objects in inferior temporal cortex** *Gabriel Kreiman*, *Chou Hung, Tomaso Poggio & James DiCarlo*, MIT McGovern Institute & BCS
10:05-10:25	**Coffee break**
10:25-10:50	**From feedforward vision to natural vision: The impact of free viewing, task, and clutter on monkey inferior temporal object representations** *James DiCarlo*, MIT McGovern Institute
10:50-11:10	**Invariant visual representations of natural images by single neurons in the human brain** *Leila Reddy [1], Rodrigo Quian Quiroga, Gabriel Kreiman, Christof Koch and Itzhak Fried*, [1] MIT McGovern Institute
11:10-11:40	**Perception of objects in natural scenes and the role of attention** *Anne Treisman, Karla Evans*, Princeton University
11:40-1:00	**Lunch break**
1:00-1:25	**Natural scene categorization: from humans to computers** *Li Fei-Fei [1], Rufin VanRullen, Asha Iyer, Christof Koch & Pietro Perona*, [1] Beckman Institute, ECE Dept, Psychology Dept, UIUC
1:25-1:50	**Contextual associations in the brain** *Moshe Bar*, *Elissa Aminoff, Nurit Gronau*, Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School
1:50-2:15	**Using the Forest to see the Trees: A computational model relating features, objects and scenes** *Antonio Torralba*, MIT Computer Science and Artificial Intelligence Lab (CSAIL)
2:15-2:25	**Coffee break**
2:25-2:45	**Detecting and remembering pictures with and without visual noise** *Mary Potter* & Ming Meng, MIT Brain & Cognitive Sciences
2:45-3:05	**Scene perception after those first few hundred milliseconds** *Jeremy Wolfe*, Brigham and Women's Hospital and Harvard Medical School
3:05-3:35	**The Artist as Neuroscientist** *Patrick Cavanagh*, Vision Sciences Lab, Department of Psychology, Harvard University
3:35-4:00	**Break**
4:00-5:00	**Brain and Cognitive Sciences Colloquium: Scene processing with a wave of spikes: Reverse engineering the visual system** *Simon Thorpe*, CNRS and SpikeNet Technology, France

9:00-9:20 From zero to gist in 200 msec: The time course of scene recognition
***Aude Oliva, Michelle Greene*, Department of Brain and Cognitive Sciences, MIT**

Human scene understanding is remarkable: with only a brief glance at an image, an abundance of information is available (spatial layout, scene function, semantic label, etc). Here we propose a scene-centered model of rapid human scene understanding that uses global scene properties as the building blocks of semantic categorization of scenes. Behaviorally, we show human observers are sensitive to the underlying distributions of these global properties, within the first 50 msec of image processing.

9:20-9:45 Feedforward theories of visual cortex predict human performance in rapid image categorization
***Thomas Serre , Tomaso Poggio,* McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, MIT**

Rapid categorization has been extensively studied over the past years. How the visual system achieves object recognition in natural images with such speed and accuracy remains however a matter of debate. Here we show that a specific implementation of a model that belongs to a class of feedforward theories of object recognition – that extend the Hubel & Wiesel simple-to-complex cell hierarchy from V1 to AIT – can predict the pattern of performance achieved by human observers on a rapid natural image categorization task.

9:45-10:05 Latency, duration and codes for objects in inferior temporal cortex
***Gabriel Kreiman, Chou Hung, Tomaso Poggio, James DiCarlo*, McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, MIT**

How much time do we need to decode information about the gist of a scene, categorize the scene or its objects and identify the individual objects present in the scene? We used a biologically plausible, classifier- based read-out technique to investigate the encoding and latency for object information in the inferior temporal (IT) cortex of macaque monkeys. IT cortex constitutes the last exclusively visual area of the ventral visual stream. We presented isolated objects while recording both spiking activity and local field potentials from a population of over 300 neurons in passively fixating monkeys using single electrodes. The activity of small neuronal populations (~100 randomly selected cells) over very short time intervals (as small as 12.5 ms) contained surprisingly accurate and robust information about both object "category" and "identity" which generalized over position and scale. Our hierarchical and feed-forward model of object recognition can account for the main observations and also make experimentally testable predictions. Information about both object identity and category could be decoded from the neuronal population starting around 100 ms after stimulus onset and saturating around 150 ms after stimulus onset.

10:25-10:50 From feedforward vision to natural vision: The impact of free viewing, task, and clutter on monkey inferior temporal object representations
***James DiCarlo*, McGovern Institute for Brain Research, BCS, MIT**

Recent work has directly demonstrated that the monkey inferior temporal cortex (IT) contains a rapidly-evoked (~150 ms), identity-explicit object representation. The key challenges going forward are an understanding of the mechanisms that underlie the IT representation and how that representation is employed to support behavior during natural vision. In particular, natural vision contains three elements that are not typically explored, but may impact the IT representation: free viewing, goal-directed behavior, and complex scenes (clutter). I will briefly present data outlining the effect of each of these elements on the IT representation.

10:50-11:10 Invariant visual representations of natural images by single neurons in the human brain.
Leila Reddy* [1], Rodrigo Quian Quiroga, Gabriel Kreiman, Christof Koch, Itzhak Fried*,**
[1] **McGovern Institute for Brain Research, MIT**

Recognition of a person, object or spatial scene can occur almost effortlessly, even when the images are viewed under remarkably different conditions. We investigated the single neuron representations of natural stimuli presented in various contexts, in the human medial temporal lobe (MTL). Beginning approximately 300 ms after stimulus onset, a subset of MTL neurons responded selectively to markedly different pictures of particular people, objects or landmarks, and in a few cases even to letter strings with their names. These responses suggest an invariant, sparse and explicit code in at least a subset of MTL neurons, which might be important in representing abstract associations between different pictures, words and other visual stimuli.

11:10-11:40 Perception of objects in natural scenes and the role of attention
***Anne Treisman, Karla K.* Evans, Department of Psychology, Princeton University**

Recent studies suggest attention-free semantic processing of natural scenes, in which concurrent tasks leave category detection unimpaired (e.g. Li et al., 2002). Could this ability reflect detection of disjunctive feature sets rather than high-level binding? Participants detected an animal target in an RSVP sequence and then reported its identity and location. They frequently failed to identify or to localize targets that they had correctly detected, suggesting that detection was based only on partial processing. Detection of targets was considerably worse in sequences that also contained humans, presumably because of shared features. When two targets were presented in RSVP, a prolonged attentional blink appeared which was almost eliminated when both targets were detected without being identified. The results suggest rapid feature analysis mediating detection, followed by attention-demanding binding for identification and localization.

1:00-1:25 Natural scene categorization: from humans to computers
***Li Fei-Fei* *1*, *Rufin VanRullen, Asha Iyer, Christof Koch, Pietro Perona*, *1* Beckman Institute, ECE Dept, Psychology Dept, UIUC**

This talk will be divided into two halves. In the first half, we will discuss psychophysics experiments that probe into the rapid perception of real-world scenes by human subjects. Our experiments show that rapid natural scene categorization can easily escape visual attention, while seemingly much simpler stimuli cannot. We will also present some recent results on experiments aimed to titrate how much humans can perceive from a briefly presented real-world image. We propose a working definition of the gist of a natural scene. In the second half of the talk, we will briefly present a computer vision model for natural scene recognition based on local patches and their textures. This probabilistic model attempts to characterize natural scenes by grouping local patches into hidden themes, and then themes into categories. We will shows experimental results of the model using a dataset of 13 classes of natural scenes.

1:25-1:50 Contextual associations in the brain
***Moshe Bar, Elissa Aminoff, Nurit Gronau*, Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School**

The parahippocampal cortex has been implicated in the processing of place-related information. It has also been implicated in episodic memory, even for abstract, non-spatial items. How could such seemingly different cognitive processes be mediated by the same cortical region? Based on our findings, we propose that the parahippocampal cortex should be viewed not as exclusively dedicated for analyzing place-related information, or as solely processing episodic memories, but instead as more generally mediating contextual associations. Contextual associations can be seen as the building blocks both for place-related information and for episodic memories; place-related information relies on associations between what and where, and episodic memory relies on associations of co-occurring entities. Therefore, by modifying the existing view of the function of this cortical region accordingly, these seemingly contradicting processes can all be explained under one overarching framework.

1:50-2:15 Using the Forest to see the Trees: A computational model relating features, objects and scenes
***Antonio Torralba,* CSAIL, MIT**

Traditional computational approaches to object detection only look at local pieces of the image. However, such local pieces can be ambiguous, especially when the object of interest is small, or imaging conditions are otherwise unfavorable. This ambiguity can be reduced by using global features of the image, which we call the gist of the scene, as an additional source of evidence. We show that by combining local and global features, we get significantly improved recognition rates. I will describe how scene information can also be used to modulate the saliency of image regions early during the visual processing in order to provide an efficient short cut for object detection and recognition. Finally, I will compare the performance of a saliency-based model and a model incorporating contextual information in predicting eye movements by participants performing a search task.

2:25-2:45 Detecting and remembering pictures with and without visual noise
***Mary C. Potter, Ming Meng*, Department of Brain and Cognitive Sciences, MIT**

Detection of a picture target in an RSVP stream, given a verbal title, is surprisingly good at rates of presentation as high as 19 pictures/s. Still more surprising, there is only a marginal drop in accuracy when 30% of the picture is covered by a random array of dots, suggesting a robust ability to fill in the picture. In contrast, memory for a picture is more markedly impaired by the dots. Rapid amodal completion may be aided by high-level information such as that given in a search title.

2:45-3:05 Scene perception after those first few hundred milliseconds
***Jeremy M Wolfe*, Brigham and Women's Hospital and Harvard Medical School**

The visual system can make meaning out of photons on the retina within a fraction of a second. Finding a specific object may take longer. Interestingly, an object, once found does not seem to have any enhanced status as a visual stimulus. If you want to find it again, you need to search again even if you have reported on the same, unchanging object many times over many minutes. Even after those many minutes, you may not notice if an object is changed or removed.

3:05-3:35 The Artist as Neuroscientist
***Patrick Cavanagh*, Vision Sciences Lab, Department of Psychology, Harvard University**

Paintings often provide compelling impressions of surfaces, light, and objects despite errors in shadows, colors, and reflections. Luckily many of these errors are not noticed by viewers and this allows artists to ignore some aspects of scene physics in order to achieve a more effective rendition. As artists discover the rules they can break without penalty, they act as research neuroscientists and we have only to look at their paintings to uncover their discoveries. The goal is not to expose the slip-ups of the masters, entertaining as that might be, but to understand the reduced set of rules that the brain uses to comprehend the world.

4:00-5:00 Brain and Cognitive Sciences Colloquium - Scene processing with a wave of spikes: Reverse engineering the visual system
***Simon Thorpe,* CNRS and SpikeNet Technology, France**

Using a saccadic eye movement task, we have recently shown that when two scenes are simultaneously flashed to the left and right of fixation, human subjects can make reliable saccades to the side where there is an animal in as little as 120-130 ms (Kirchner & Thorpe, 2006, Vision Res). Such results imply that complex decisions can be made under conditions where not only do individual neurons not have time to emit more than one spike, but where the proportion of neurons that will have fired in any particular visual structure may be very low. Simulations with SpikeNet, an image processing system based on large-scale networks of asynchronously firing neurons show that such an approach is indeed feasible. We have found that when the ordering of firing is taken into account, neurons in higher order visual areas can be trained to respond to the presence of particular diagnostic features (such as an eye, a beak and so on) very efficiently. Furthermore, resistance to variations in luminance and contrast as well as blurring and noise is remarkably good. Invariance is also good to rotation (around 10), changes in size (15%) and angle of viewing. We propose that with a very large number of such mechanisms trained to detect particular diagnostic feature combinations, performance in ultra-rapid categorization tasks could be explained by supposing that reliable behavioral responses are initiated when a sufficient number of these diagnostic feature combinations has been detected. Finally, we have observed that a biologically plausible learning algorithm based on Spike Time Dependent Plasticity will naturally lead to a situation where neurons become selective to frequently encountered visual patterns by concentrating high synaptic weights on the earliest firing inputs.

SUNS 06 is sponsored by the Center for Biological & Computational Learning, the Department of Brain and Cognitive Sciences, and the McGovern Institute for Brain Research, MIT

Acknowledgments: Many Thanks to Winston Chang, Emily Connally, Barbara Hidalgo-Sotelo, Talia Konkle, Elisa McDaniel, Bettiann McKay.