NSF Workshop on Hybrid Neuro-Computer Vision Systems
April 19-20, 2010 || Columbia University, New York City, NY, USA

Workshop URL: http://www.columbia.edu/cu/hybridvision/

DAY 1: APRIL 19, 2010
Interschool Lab, CEPSR 7th floor

8:45 Introduction

9:00 -10:40 (AREA 1: Neuroscience and Neural Computing)

Neural Mechanisms of Feature Selectivity in Natural Vision: Detection and Discrimination
Garrett Stanley, Georgia Institute of Technology

Despite decades of research in systems neuroscience, there still exists a fundamental debate over the basis of the neural code and, specifically, what aspects of the neural activity are important for neural computation. Our laboratory has focused on this problem in the context of neural representations of different features in the natural visual world. Through population activity in the visual thalamus, we have shown that the temporal precision in the neural response changes relative to the timescale of the visual scene, that a high degree of temporal synchrony is necessary to represent the more slowly varying natural scene, and that the ensemble representation is invariant to visual contrast. More recently, we have shown that synchronous ensemble activity expresses a sharp tuning for direction of motion in the visual scene, beyond that predicted by conventional linear models, and can be used to decode motion direction and speed over short timescales on a single trial basis. Finally, recent simultaneous recording of thalamus and cortex has revealed a contextual dependence of neural activity for detecting and discriminating between different features of the sensory input, suggesting a dynamic relationship between the neural activity and the features of the sensory environment that are being represented in the thalamocortical circuit. Taken together, the work from our laboratory suggests an explicit role for modulation of timing across neural populations for computation in the natural sensory world.

Network Models of Sparse Coding and Nonclassical Receptive Field Effects
Chris Rozell, Georgia Institute of Technology

Sparse coding theories postulate that V1 should encode images compactly by using relatively few co-active units. While this encoding theory has successfully predicted the characteristic shapes of measured V1 receptive fields, the actual encoding requires the neural population to solve a relatively complex optimization problem. I will present a network model, specified as a dynamical system with neurally plausible computational primitives, that provably solves the optimization problems required for sparse coding. I will also present results from simulated physiology experiments showing that this model can demonstrate several non-classical receptive field effects, including end-stopping, surround suppression, and contrast invariant orientation tuning.

Learning Spatial and Transformational Invariants for Visual Representation
Charles Cadieu, University of California Berkeley

I will describe a hierarchical, probabilistic model that learns to extract complex spatial structure and complex motion from movies of the natural environment. The first layer in the model produces a sparse decomposition of the local edge and motion structure. This

decomposition exposes statistical dependencies that the top layer in the model captures as two types of invariances: spatial and transformational. The spatial invariants encode geometric structure and texture characteristics; and the transformational invariants encode the way objects move irrespective of the way they look. The hierarchical model provides a concrete model of cortical feedback that I will show is useful for perception under noisy or ambiguous input conditions.

Sparse Decoding of Neural Dynamics Generated by Large Scale Models of Primary Visual Cortex
Paul Sajda, Columbia University

In this talk I will describe our work investigating sparse decoding of neural activity, given a realistic mapping of the visual scene to neuronal spike trains generated by a model of primary visual cortex (V1). We use a linear decoder which imposes sparsity via an L1 norm. The decoder can be viewed as a decoding neuron (linear summation followed by a sigmoidal nonlinearity) in which there are relatively few non-zero synaptic weights. We find: (1) the best decoding performance is for a representation that is sparse in both space and time, (2) decoding of a temporal code results in better performance than a rate code and is also a better fit to the psychophysical data, (3) the number of neurons required for decoding increases monotonically as signal-to-noise in the stimulus decreases, with as little as 1% of the neurons required for decoding at the highest signal-to-noise levels, and (4) sparse decoding results in a more accurate decoding of the stimulus and is a better fit to psychophysical performance than a distributed decoding, for example one imposed by an L2 norm. We conclude that sparse coding is well-justified from a decoding perspective in that it results in a minimum number of neurons and maximum accuracy when sparse representations can be decoded from the neural dynamics.

10:40-10:55 BREAK

10:55-12:10 (AREA 1: Neuroscience and Neural Computing)

Voxel-based encoding models in fMRI and their use for decoding
Jack Gallant, University of California, Berkeley

Most current work on brain-computer interfaces focuses on decoding algorithms that aim to classify, identify or reconstruct the stimulus directly from measured brain activity. These approaches are often based on non-parametric algorithms such as support vector classification that do not include any explicit information about how the brain encodes information. My laboratory has pioneered an alternative approach that is based on estimating explicit non-linear encoding models that describe how stimuli are transformed into measured brain activity. These models can be evaluated in the forward direction to investigate how the brain represents visual information. Furthermore, they can be inverted and combined with an appropriate prior to obtain Bayesian decoding models that can be used to reconstruct perceptual experiences. This Bayesian decoding framework is quite general: in theory it can be applied to any brain activity measurements, and can be used to reconstruct many perceptual and cognitive states. Thus, Bayesian decoding algorithms might form the basis of powerful new brain-reading technologies and brain-computer interfaces.

Decoding Visual and Mental States from Human Brain Activity
Frank Tong, Vanderbilt University

Our research has revealed that considerable information about visual and mental

states can be decoded from patterns of activity in the human visual cortex. Information that primarily resides at fine spatial scales, such as orientation selectivity, can be decoded from coarse-scale activity patterns due to local variability in the fine-scale cortical map. More striking, decoding can be applied to read out subjective mental states, and can reliably predict what visual feature or object is being attended by an observer, or what item is being maintained in working memory. Decoding therefore provides a powerful new method to investigate how early visual areas support diverse cognitive functions of perception, attention, and working memory.

Estimating Dynamics of Information Processing from EEG and MEG Measurements
Philippe Schyns, Glasgow University

To study visual categorization mechanisms in an information system such as the brain, three generic questions must be addressed to relate brain activity to the states of an automaton (i.e. to derive an algorithm). The first question is that of form: What is the nature of the brain activity supporting face processing (i.e. the states of the brain correlated with face processing?) The second question is that of content: What is the information content processed in these brain states? The third question is that of transition: How does information flow from one brain state to the next between stimulus onset and behavioural response? I will present new methods and results that illustrate how information states can be estimated and interpreted directly from brain activity.

12:10-1:30 LUNCH, CEPSR 414

1:30-2:45 (AREA 2: Neuro-Inspired Computer Vision)

Bottom-up and top-down processing in visual perception
Thomas Serre, Brown University

Perception involves a complex interaction between feedforward sensory-driven information and feedback attentional, memory, and executive processes that modulate such feedforward processing. A mechanistic understanding of feedforward and feedback integration is a necessary step towards elucidating key aspects of visual and cognitive functions and dysfunctions. In this talk, I will describe a computational framework for the study of visual perception. I will present computational as well as experimental evidence suggesting that bottom-up and top-down processes make a distinct and essential contribution to the recognition of complex visual scenes. A feedforward hierarchical architecture may provide a satisfactory account of “immediate recognition” corresponding to the first few hundred milliseconds of visual processing. However, such an architecture may be limited in recognizing complex visual scenes. I will show how attentional mechanisms and cortical feedback may help improve object recognition performance in complex cluttered scenes.

Learning Deep Hierarchies of Visual Features
Yann LeCun, New York University

The visual cortex is a deep architecture in which the successive stages produce representations of the visual world that are increasingly global, invariant, and

abstract. How can the visual cortex build this hierarchy of internal representation by merely looking at the world? Can we devise learning algorithms that would automatically learn such a hierarchy of representations from unlabelled data? I will describe a class of "deep learning algorithms" that can learn feature hierarchies in an unsupervised manner. The features can be uses to learn visual categories with a very small amount of labeled data. A real-time demo of an object recognition system will be shown. An application to mobile robots that learn vision-based navigation autonomously will be described.

Visual Scene Understanding
Aude Oliva, Massachusetts Institute of Technology

Visual scene understanding is central to our interactions with the world. Recognizing the current environment facilitates our ability to act strategically, for example in selecting a route for walking or anticipating where objects are likely to appear. Converging evidence from behavioral, computational and cognitive neuroscience studies suggests that the human brain may employ a strategy for representing the gist or meaning of scenes that is independent of the visual complexity of the image, and that, under specific conditions, human observers have a phenomenal long-term memory capacity for the visual details of complex scenes. Together, these remarkable feats of human perception and memory suggest new avenues of research for understanding and modeling the mechanisms that underlie real-world scene recognition.

2:45-3:00 BREAK

3:00-4:15 (AREA 2: Neuro-Inspired Computer Vision)

What, where and whom: what do we see in a glance of a scene? And what can computers see?
Li Fei-Fei, Stanford University

Given a glance of a real-world picture, humans can recognize the objects, locate their position, name the overall scenery, describe the event and/or social activity, and even tell the mood of the players and atmosphere of the event (Fei-Fei et al. J. of Vision, 2007). Can we make computers do the same? More importantly, inspired by the large amount of data available through the Internet, however noisy they are, can we develop an automatic learning framework to allow algorithm to learn these complex concepts without much human supervision? In this talk, I present a number of recent work in our lab towards understanding real-world objects, scenes and activities involving human-object interactions (Li & Fei-Fei ICCV, 2007; Li et al. CVPR, 2009, 2010; Yao & Fei-Fei, CVPR, 2010a, b).

Large-scale Scene and Object Recognition
Antonio Torralba, Massachusetts Institute of Technology

I will present results on scene and object recognition obtained with a new database of more than 400 scene categories and more than 100 object classes. When hundreds of categories become available new challenges and opportunities emerge. One of the challenges is to devise efficient and accurate algorithms able to handle hundreds and thousands of categories. But there are also new opportunities. On

scene recognition, we can test the performance of global features to classify scenes into a very large number of possible settings covering most of the places encountered by humans. On the object recognition side, we can develop context based models that will benefit from the interactions between hundreds of object categories. For instance, the performance benefit from integrating context models into object detection has been limited because most of these methods were tested on datasets with only a few object categories, in which most images contain only one or two object categories. In this work, we introduce a new dataset with images that contain many instances of different object categories and propose an efficient model that captures the contextual information among more than a hundred of object categories.

Biologically-inspired Vision and Attention for Cognitive Robots
Laurent Itti, University of Southern California

Recent years have witnessed tremendous advances in endowing robots with autonomous reasoning and decision-making capabilities. This has given rise to highly intelligent and cognitively capable machines, which in some cases approach or exceed human abilities. However, one aspect in which robots are still lacking compared to their biological counterparts is in their sensory and motor interaction with the real world, including: rapidly finding and identifying objects of interest in cluttered scenes, building cognitive representations of scenes, and physically interacting with these scenes. Here, I will review a number of exciting new algorithms which draw inspiration from biology to attempt to bridge the gap between artificial and natural visual systems. Specifically, I will describe:
- key components of primate vision, including bottom-up and top-down attention, object recognition, rapid computation of the 'gist' of a scene, and visual cognitive reasoning;
- supporting evidence for a simple yet highly flexible architecture which orchestrates the operation of these modules so as to support a wide range of visual cognitive tasks;
- examples of successful robotics and machine vision systems which have used the above computational principles and modules, and have demonstrated strong performance at object detection in cluttered scenes, scene parsing, robot localization in outdoor natural environments, and accurate prediction of where humans look when searching for particular items in complex scenery.

4:15-4:30 BREAK

4:30-5:45 (AREA 3: Hybrid Vision Systems and Applications)

Why Neuro? – A Perspective on Keeping the Human in the Loop
Amy Kruse, Total Immersion Software

We have crossed a threshold. This past year the Air Force graduated more UAV pilots than pilots to fly manned airframes. In 2009, the number of UAV combat flight hours exceeded 200,000. That number will only increase. The Unmanned Air Vehicle (UAV) is a bit of a misnomer – although the human is not in the airframe – they are certainly “in the loop”. The use of unmanned sensors, whether air or ground, will continue to increase exponentially, as it has over the past decade. Sensors are being deployed (Reaper, Gorgon Stare, ARGUS) with 10, 30

and eventually 65 independent feeds – with a human on the receiving end of that information. Now, more than ever, there is a pressing need for smart sensors – and smart sensor systems that combine the "best of" processing from computational systems and humans. This talk will discuss the current battlefield needs and highlight the importance of hybrid neuro-vision systems.

Neuro-inspired Statistical Prior Models for Computer Vision
Qiang Ji, Rensselaer Polytechnic Institute and NSF

Many computer vision problems are ill-posed. Data-driven approaches, no matter how advanced they are, will likely to fail under many realistic conditions. On the other hand, human perception can perform many challenging vision tasks with ease. One factor contributing to this is due to human's capability of capturing related prior knowledge and integrating the prior knowledge with the visual measurements for robust visual inference.

To emulate this human ability, we propose a statistical framework based on the probabilistic graphical models (PGMs) to capture prior knowledge from various sources. Recent studies show that PGMs such as Bayesian networks, MRF, and factor graph match well with structural and algorithmic properties of the neocortex in information representation and processing. In fact, PGMs are the basis of several well known neuro-inspired computational models.

For this research, we introduce advanced machine learning methods to construct PGM prior models from different sources. The knowledge sources include the underlying statistical properties of the image, the natural and inherent relationships among different image entities, and the geometric, anatomical, physical, and dynamic knowledge that govern the behaviors of the objects being studied. Given the models and image measurements, visual understanding can be formulated as a probabilistic inference problem. In this talk, I will discuss how we construct the prior models using knowledge from different sources for three vision applications: image segmentation, facial action recognition, and human body tracking.

Brain State Decoding for Rapid Image Retrieval
Shih-Fu Chang, Columbia University

Human vision system is able to recognize a wide range of targets under challenging conditions, but has limited throughput. Machine vision and automatic content analytics can process images at a high speed, but suffers from inadequate recognition accuracy for general target classes. In this talk, we present a new paradigm to explore and combine the strengths of both systems. A single trial EEG-based brain machine interface (BCI) subsystem is used to detect objects of interest of arbitrary classes from an initial subset of images. The EEG detection outcomes are used as noisy labels to a graph-based semi-supervised learning subsystem to refine and propagate the labels to retrieve relevant images from a much larger pool. The combined strategy is unique in its generality, robustness, and high throughput. It has great potential for advancing the state of the art in media retrieval applications. We will discuss the performance gains of the proposed hybrid system with multiple and diverse image classes over several data sets, including those commonly used in object recognition (Caltech 101) and remote sensing images. (joint work with J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, and P. Sajda)

6:30-9:30 DINNER
Terrace in the Sky, 119th St and Morningside Dr.

DAY 2 APRIL 20, 2010

9-12 Breakout Sessions (2 groups)
CEPSR 7th floor Interschool Lab and CEPSR 414

12-1 LUNCH, CEPSR 414

1-3 Plenary Session: group report and discussion
CEPSR 7th floor Interschool Lab