Abstracts of NIH-Conte Meeting – August 30, 2004

Position-Specific Tuning for Boundary Conformation in the Standard Model of Object Recognition

Charles Cadieu, Tomaso Poggio
Massachusetts Institute of Technology

The computational processes in the intermediate stages of the ventral pathway for visual object recognition are not well understood. A recent physiological study by A. Pasupathy and C. Connor in intermediate area V4 indicates that neuron responses display object-centered position-specific curvature tuning (A. Pasupathy C. Connor, 2001). The Standard Model of object recognition, a recent model for biological object recognition developed by M. Riesenhuber and T. Poggio, does not have explicit representation for either object center or curvature. However, novel model units constructed within the framework of the Standard Model exhibit response patterns similar to V4 cells. The response of these units displays high correspondence to raw responses of V4 cells and also shows similar tuning in the shape spaces developed by A. Pasupathy and C. Connor. These results suggest an alternative interpretation of V4 cell tuning that is not based on object center or curvature. V4 cell tuning may be more appropriately described as tuning to Gabor filters at specific relative positions.

Shape Tuning in Monkey Inferior Temporal Cortex (ITC) is Enhanced by Experience

David Freedman, Maximilian Riesenhuber, Tomaso Poggio, Earl Miller
Massachusetts Institute of Technology

We had trained monkeys to categorize computer-generated stimuli into two categories ("cats" and "dogs") and found that ITC neurons showed robust stimulus-selectivity. The stimuli always appeared at the same orientation during weeks of training. Here, we tested whether this experience resulted in enhanced selectivity at the trained orientation compared to six untrained image-plane rotations. We also (using other stimuli) asked whether enhanced tuning was due to explicit training or if it could result from passive experience.

We recorded from 186 ITC neurons while two alert monkeys passively viewed 18 cat and dog stimuli at the trained and six rotated orientations (22.5, 45, 67.5, 90, 135 and 180 deg.). We also recorded from 298 ITC neurons while two monkeys viewed highly familiar (n=20) and novel (n=20) randomly-chosen, complex stimuli.

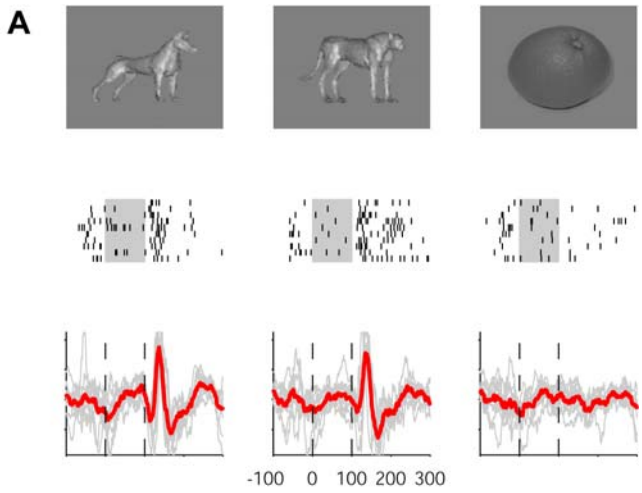
Average firing rates were not greater for trained/familiar stimuli than untrained/novel stimuli. However, ITC neurons were more sharply tuned to the cat and dog stimuli at the trained vs. untrained orientations. Similarly, tuning was sharper for the familiar, passive-experienced stimuli relative to novel stimuli. Sharpened tuning for trained/familiar stimuli was primarily evident early in the visual response (80-180 ms post-stimulus). This suggests that ITC shape tuning is shaped by visual experience.

Selectivity and Clustering of Spikes vs. LFPs in Macaque Inferotemporal Cortex

Chou Hung, Gabriel Kreiman, Tommy Poggio, Jim DiCarlo
Massachusetts Institute of Technology

It is well known that neural spiking activity in the inferotemporal cortex can be very selective for complex pictures. We investigated whether local field potentials (LFPs), which are thought to reflect activity from large numbers of neurons, also exhibit such selectivity. Using single electrodes, we recorded multi-unit activity (MUA) and LFPs from the anterior inferior temporal (AIT) cortex of macaques while they passively viewed 77 different pictures of complex stimuli. Strikingly, the LFP data also showed statistically significant selective responses. Typically, the LFP response was selective to less than 10% of the presented pictures and showed less selectivity than MUA data. Control experiments from shuffled data and also from recordings outside the inferior temporal cortex did not show selectivity. Surprisingly, our preliminary observations suggest that there was no strong overlap between the selectivity of MUA and LFP recordings from the same electrode.

Selectivity from MUA activity at a given site was correlated with the selectivity observed at a site separated by approximately 200-800 μm. At longer distances (> 1mm), correlations were higher for LFPs than for MUA selectivity. This observation, together with the LFP data suggests that there is some topographical arrangement to the organization of selectivity in inferior temporal cortex.



Testing Model Predictions in Psychophysics and fMRI

Xiong Jiang, Maximilian Riesenhuber
Georgetown University

In our current work, we are exploring whether our computational model of object recognition in cortex can be used to derive quantitative predictions for human psychophysics and fMRI, thereby bridging the different levels of description of the visual system, from single neurons to behavior. Two pilot studies (one fMRI, one psychophysics) were conducted to test model predictions. In the first study, an fMRI adaptation paradigm was used to address two questions: (1) Whether the model could be used to make quantitative predictions for fMRI experiments, and (2) whether face processing in cortex could be well described by the model (a question of special interest since many existing theories of human face recognition claim that face recognition is "special", in contrast to physiological data). In an fMRI rapid adaptation paradigm, two visual stimuli are presented in rapid succession. Physiological studies have established that the level of response to the second stimulus in the pair is a function of its similarity to the first stimulus, with increasing levels of response suppression occurring for more similar stimuli. We used the model to derive predictions for the amount of adaptation of "face neurons" in the human FFA (Fusiform Face Area) as a function of intra-pair similarity for image pairs consisting of two faces. Four different degrees of similarity (generated using the morphing system by Blanz & Vetter) between the two successively displayed faces were used. A strong quantitative coincidence between the model predictions and fMRI data was found for three pilot subjects, suggesting that FFA neurons can be modeled as shape-tuned units, and that the model might be helpful to design further quantitative fMRI studies. The psychophysical study was conducted to test the model prediction of a U-shaped interaction for target detectability as a function of dissimilarity between target and simultaneously presented distractor. In contrast to our earlier studies where we had trained human subjects to become experts for a novel object class (morphed cars), we are now exploring whether the effect can be obtained for faces, an object class human subjects are already experts for (thus eliminating the need for training and further testing the neural mechanisms underlying the cognitively important class of faces). While this study will ultimately require a subject-specific calibration of face morph space to reflect subjects' perceptual similarity space, pilot data without this adjustment already found a U-shaped interaction for two out of three subjects. More data need to be gathered, but these initial results provide an encouraging basis for future studies to use the model as a framework to also study human cognition.

A Hypothetical Ferster Circuit for the MAX Operation in V1

Ulf Knoblich, Tomaso Poggio
Massachusetts Institute of Technology

The MAX function has been postulated to play an important role in object recognition in visual cortex [Riesenhuber & Poggio 1999]. It is thought to be essential in order to achieve selectivity and invariance found in complex cells in visual cortex. Recently, evidence for the presence of single cells which compute a soft-MAX operation has been found in cat primary visual cortex [Lampl *et al.* 004)].

We investigated a model of a synaptic mechanism to perform this computation utilizing both excitatory and inhibitory synaptic conductances for each input. We assume that each input gives rise to an excitatory as well as an inhibitory postsynaptic potential which is created through an interneuron. Due to the nonlinear response properties of the interneuron, the ratio of excitatory to inhibitory input increases with increasing input. This causes the membrane potential to equilibrate at different levels at the post-synaptic site. By combining two of these inputs, a membrane potential results that closely resembles the potential induced by the larger of the two inputs (“soft-MAX") because of the local interaction of the excitatory and inhibitory inputs.

This model provides a good fit for the intracellular measurements in a population of complex cells found in primate V4, cat primary visual cortex and rat barrel cortex. According to the model, interaction occurs at the synapse but the cell's soma does not saturate. This prediction should be easily testable and will guide future experiments.

V1 and V4 in the Model

Minjoon Kouh, Tomaso Poggio
Massachusetts Institute of Technology

We show that the bowtie-like pattern in the two-spot reverse correlation experiment (Livingstone and Conway, 2003) can be explained by assuming that the neural response is normalized by the total response from a pool of other neurons, including itself. The Riesenhuber-Poggio model still works (in terms of its VTU tuning to the paperclip stimuli) with this particular form of normalization at the C1-layer. We note that the normalization, along with the weighted sum, could be the underlying general mechanism for cortical tuning. Furthermore, the neural circuitry for another key operation in the model, maximum operation, may be quite similar to the circuitry for the normalization.

Selectivity of Local Field Potentials in the Human Brain

Alexander Kraskov, Rodrigo Quian Quiroga, Christof Koch, Itzhak Fried
California Institute of Technology

We started to investigate local field potentials in the human temporal medial lobe. The recordings are done from the brain of epileptic patients in collaboration with Dr. Itzhak Fried, UCLA. The main goal of this project is to characterize the selectivity of the LFP’s to different image categories (e.g., photos of faces versus pictures of landmark buildings) or to individual images and to study the relationship between the selectivity of single and multi units and the LFPs. That is, if we record from a neuron that shows great selectivity of faces or even to the face of a single individual, will the LFP recorded from the same electrode also show such selectivity or a different one? This would be important for practical purposes since recording LFPs is much easier and more robust to electrode movement that picking up spiking activity from a single neuron. Stimulus selectivity in the LFP would also reveal spatial clustering of neurons, that is, the fact that neurons with similar stimulus selectivity (e.g., to faces) cluster in space.

We are interested in extracting specific information from the LFPs and therefore we choose continuous wavelet transform method to explore time and frequency structure of LPFs. This method allows us to consider selectivity in individual frequency bands (standard electroencephalographic frequency bands δθααβγ) and in different time intervals (well known peaks of evoked responses).

We started from the analysis of power spectrum that looks quite smooth and has and 1/f appearance with average α about 1.9. Moreover, in many electrodes we found significant oscillations in different frequency bands, e.g., θα). As a preliminary result we found selectivity to animal category in the electrodes that were also selective according to single and multi units.

Decoding Neural Signals from IT

Gabriel Kreiman, Chou Hung, Tommy Poggio, Jim DiCarlo
Massachusetts Institute of Technology

How accurately can we read out information about objects from neuronal activity in the brain? To attempt to decode the information from spikes and LFP data, we applied a classifier to discriminate between 8 arbitrarily pre-defined groups of stimuli drawn from a set of 77 images. Assuming independence between different sites, we estimated how classification performance changes with the number of sites, and other parameters including the bin size and integration time. Preliminary results suggest that an SVM classifier performs better than a Fisher linear discriminant classifier, and that performance increases linearly with the log of the number of sites. Classification performance is better for spiking activity than classification based on LFPs. Future work will explore the invariance properties of the neural activity to changes in size, position, and illumination.

The Attentional Requirements of Face Processing

Leila Reddy
California Institute of Technology

The attentional cost associated with the visual discrimination of the gender of a face was investigated. Participants performed a face-gender discrimination task either alone (single-task), or concurrently (dual-task) with a known attentional demanding task (5-letter T/L discrimination). Overall performance on face-gender discrimination suffered remarkably little under the dual-task condition compared to the single-task condition. Similar results were obtained in experiments which controlled for potential training effects, the use of low-level cues in this discrimination task or eye movements.

Further experiments investigating the attentional cost of face recognition demonstrated that under the same conditions as above, subjects were able to discriminate between faces at the level of the individual when attention was removed. Thus, our results provide further evidence against the notion that only low-level representations can be accessed outside the focus of attention.

How a Cortical Model Learns a Vocabulary of Visual Features from Images

Thomas Serre, Rodrigo Sigala, Robert Liu and Tomaso Poggio
Massachusetts Institute of Technology

Learning in the standard model was so far limited to the highest layers and all synaptic weights in lower and intermediate layers were 'static'. We previously showed that this simple architecture was sufficient for recognition of idealized stimuli (paperclips on blank background) but inadequate for real-world object recognition such as face-detection in clutter [Serre *et al.*, 2002].

We present a long-planned model extension for learning a dictionary of visual features from image sequences. Our learning rule relies on a trace mechanism (record of the recent synaptic activity), which is both simple (hebbian) and locale (no distributed back-propagation). Extending previous work that has shown that a simple trace rule could allow a network to learn shift invariance [Foldiak 91], our results suggest that the same simple mechanism could also be used in cortex to learn other kind of invariances such as pose, illumination and intra-class variations (between individuals). After introducing a stage of features learning our biologically motivated model of object recognition performs surprisingly well – better than state-of-the art computer vision algorithms – on a range of recognition problems. Our model results suggest looking for plasticity in visual cortex.

Psychophysics of Object Competition

Rufin VanRullen, Christof Koch
California Institute of Technology

Binding is often referred to as the process by which basic features of an object are conjoined within the focus of attention to allow recognition. We have previously argued, however, that certain high-level objects can be recognized outside the focus of attention. We proposed that binding exists in fact under two forms. The visual system heavily relies on hardwired binding whereby relevant objects and feature conjunctions are selectively coded by dedicated neuronal populations (e.g. faces, animals, color-orientation conjunctions). Attention is not required for this form of binding, but must sometimes be engaged to resolve spatial competition within a receptive field. Without such hardwired selectivities, binding of arbitrary feature conjunctions can still occur but necessitates attention (e.g. bisected 2-color disks, randomly rotated letters). Here we verify that the first postulated form of binding (hardwired) is indeed pre-attentive, but only parallel when stimuli are reasonably separated: under dual-task conditions where full attention is unavailable, a single peripheral animal, face or color-orientation conjunction can be recognized, but will suffer from the addition of a second distracting stimulus (natural scene, face, etc.), only if it is placed in the vicinity of the target stimulus. In other words, the hardwired binding problem might be receptive-field specific. This is confirmed by a second experiment in which two simultaneous masked stimuli must be compared. The SOA is chosen so that a similar stimulus in isolation is easily identified. With animal vs. non-animal scenes, human faces or color-orientation conjunctions, this comparison task can be performed with well-separated but not with neighboring stimuli. In contrast, bisected disks or randomly rotated letters cannot be compared even at large spatial separations. Thus, while arbitrary binding always requires attention, hardwired binding only does when receptive field competition occurs.

On the Usefulness of Selective Visual Attention for Object Recognition

Dirk Walther, Christof Koch
California Institute of Technology

A key problem in learning representations of multiple objects from unlabeled images is that it is a priori impossible to tell which part of the image corresponds to each individual object, and which part is irrelevant clutter that is not associated with the objects. Clutter hurts object recognition, because it generates false alarms and imposes additional computational costs for rejecting them. Distinguishing individual objects in a scene would allow unsupervised learning of multiple objects from unlabeled images. There is psychophysical and neurophysiological evidence that the brain, which is faced with a similar challenge, employs selective visual attention to select relevant parts of the image and to serialize the perception of individual objects. We propose a method for the selection of salient regions likely to contain objects, based on bottom-up visual attention, in order to allow unsupervised one-shot learning of multiple objects in cluttered images. By comparing the performance of David Lowe's recognition algorithm with and without attention, we demonstrate in our experiments that the proposed approach can indeed enable learning of multiple objects from complex scenes, and that it can strongly improve learning and recognition performance in the presence of large amounts of clutter.

In addition to bottom-up attention, top-down attention would enable biasing visual perception in accordance with prior knowledge about a particular scene or situation, or with a particular task. This task could be, for instance, visual search in natural scenes. For such scenarios, mapping a task to a particular set of biases of the visual system is a notorious problem. In our standard model of object recognition in cortex (see Thomas Serre's work in the Center), we learn a set of complex features that are suitable for the detection of particular classes of objects (e.g. faces). We suggest to use this learned set of features and the existing association with object classes for biasing visual perception in a top-down, task-dependent fashion. We have conducted first successful experiments with using skin color as a top-down bias for face detection, and we are currently assessing the use of the learned S2 features in the standard model for top-down biasing.

Neural Mechanisms underlying Visual Object Recognition in Clutter

Davide Zoccolan, James DiCarlo
Massachusetts Institute of Technology

The first step in our experimental design was to train monkey subjects to be experts at detecting specific objects. One monkey subject has been trained to become expert in an object recognition task that requires the detection of a specific shape (the *target shape*) embedded in a temporal sequence of shapes drawn from the same, parameterized shape space (the *distractors*). To insure the generality of our results, the monkey has been trained to detect a target object in each of three different parameterized shape spaces (cars, faces, and abstract silhouettes). The results of the training in each of these spaces showed a consistent performance improvement (more than doubling) during the first 7-10 days of training that reached an asymptotic value that remained constant for the remaining 8-10 training sessions.

Once the behavioral training has been completed, we started to perform single unit recordings from IT cortex. Each isolated neuron is tested for: 1) responsiveness of the neuron to stimuli sampled from the trained spaces; 2) selectivity of the neural response across the optimal stimulus space; 3) position tolerance of the shape selectivity; 4) impact of clutter on the shape selectivity. All the recordings are performed in passive viewing rapid sequence presentation (5 stimuli per second).

Our preliminary recordings are encouraging in many regards. We found neurons sharply tuned across subsets of shapes belonging to our stimulus spaces. The response of these neurons is maximal for a specific shape (the *optimal* stimulus) and then smoothly decreases for stimuli more and more dissimilar from the optimal one. We were also able to test the interference produced by clutter, i.e. pairs of stimuli of controlled similarity simultaneously present in the neuron's receptive field. These first recordings showed that the neuronal response smoothly decreases as a function of the distance, in the shape space, of the flanker stimulus from the optimal one. Our plan is to keep recording IT neuronal responses in this monkey for at least the next four months in order to obtain a quantitative measurement of clutter interference and impact of learning in IT, which are the main goals of our current project.