MIT Progress Report Summary - 2005

1. Specific Aims

The MIT project involves both computation as well as physiology in monkeys. Computational modeling of visual cortex may interact daily with experiments in the nearby physiology labs of E. Miller and J. DiCarlo. We continue to be guided by the model of recognition, which itself evolving as an effect of the experiments, in our efforts to understand the properties of selectivity and invariance of recognition, especially with image clutter, in IT and PFC cortex of behaving macaque monkeys and the relations between identification and categorization.

In particular, DiCarlo and Poggio (with Riesenhuber) are testing the effects of clutter on the selectivity and invariance of IT neurons. In the second part of this project Miller and Poggio (with Riesenhuber) are investigating the neural bases of the recognition tasks of *identification and categorization.* In addition, Poggio (with Koch) is working on biophysically plausible circuits for the two key operations in the recognition model – the max-like operation and the Gaussian-like multidimensional tuning. The work involves a close collaboration with CalTech on the computational side and with Northwestern on the experimental side.

Our specific aims are listed below from the original proposal:

Aim A.1: To determine the baseline IT neuronal relationships of a) shape-selectivity and clutter-tolerance, and b) position-tolerance and clutter-tolerance.

Aim A.2: To re-examine the relationship of shape-selectivity and clutter-tolerance and the relationship of position-tolerance and clutter-tolerance in the same monkeys after extensive training in clutter.

Aim B.1: To determine if there is a common neural substrate for different recognition tasks.

Aim B.2: To study the neural bases of the interaction of identification and categorization in Categorical Perception.

Since the beginning of our project, we added three new collaborative aims:

Aim N.1: To explore the mechanisms underlying the Gaussian-like tuning of cortical cells.

Aim N.2: To refine the model and check its prediction about properties of V4 cells (using data from Reynolds and Desimone and especially in ongoing collaborations with the Harvard lab of Dr. Livingstone and the JHU lab of Dr. Connors).

Aim N.3: To test the performance of the model on complex, natural images and compare it to the performance of humans and the response of neurons in the monkey IT and PFC.

2. Studies and Results

"New" aim N.1: We planned to study computationally (with CalTech) and experimentally (with Northwestern) the circuitry underlying the max operation. A paper has been published in the meantime. While we are continuing to work on this problem, we also started to work on the related and new question of the neural basis for the other key operation in the model: Gaussian-like multidimensional tuning of cortical cells. We are focusing on a Gaussian-like operation described as a normalized dot product, followed by a sigmoidal nonlinearity. The operation may be implemented by a local circuit utilizing shunting inhibition.

"New" aim N.2: We are investigating the feature selectivity within the intermediate visual areas before IT (such as V4 and V2), by replicating the tuning properties of the neurons to various visual stimuli. While incorporating more realistic normalization-based Gaussian-like tuning (Kouh and Poggio, 2004) in the model, we have developed a systematic methodology to quantitatively fit shape selectivity of model neurons to experimental data (Cadieu *et al*. 2005). Using this technique, we have shown that the position-specific tuning for boundary conformation of V4 neurons (Pasupathy and Connor 2001) is consistent with the model. We are also comparing and analyzing the responses of V4 and model neurons to gratings and sparse noise stimuli, using the data from Freiwald and Livingstone (2005). Furthermore, we can fit well the data of Reynolds, Chelazza and Desimone on a weighed average effect in V4 (in the absence of attention). With the currently available data, the estimation of model parameters is underconstrained and produces multiple solutions but it is still possible to obtain a set of model units with consistent shape tuning and invariance properties. We can use the set of candidate model units to make predictions to a new set of stimuli. We are continuing to collaborate with V4 physiologists (Connor, Freiwald, and Livingstone) on both the modeling of existing data and the design of new experiments to gain further understanding of intermediate shape tuning in the ventral stream.

"New" aim N.3: We recently proposed a new biologically plausible learning rule. When passively exposed to many natural images the model generates a dictionary of shape-tuned units from V4 to IT. We found that learning improves drastically the recognition performance of the model in clutter (aim A.2). Extensive tests on large-scale real-world object-recognition tasks suggest that the model can: 1) handle the recognition of many different object categories in clutter, 2) learn from very few training examples (on the order of 30 positive training examples) and 3) outperform some of the best computer vision systems. Preliminary results (see new aim N.2) already suggest that the tuning properties of the C2 units generated by the model after learning are consistent with V4 data (Pasupathy and Connors, 2001). The ability of the model to generate a set of shape-tuned units compatible with neural data and at the same time compete with state-of-the-art computer vision systems suggests that we have the capability to make robust non-trivial predictions. We plan to compare the model performance with both humans and a population of IT neurons on various difficult -- but feedforward -- recognition tasks (ultra-rapid categorization e.g. animal *vs.* non-animal, face *vs.* non-face, *etc*). We conjecture (see CalTech report) that computationally difficult tasks such as categorization of natural scenes (e.g. as containing animals and non-animals) can be done in a feed-forward architecture. If simulations of the standard model and human psychophysics will support the conjecture, we may then be able to argue that we have a satisfactory model of the feedforward path in the ventral stream of visual cortex.

Aim A.1: We have made significant progress on this aim. Three monkeys have been trained in object recognition tasks using isolated visual stimuli. Two monkeys have been trained to identify specific visual objects with different eye movements and one monkey has been trained in a *sequential 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*). We have collected response data from single IT neurons in all three of those monkeys (104 IT neurons to date). The main visual conditions we have focused on thus far are rapid, passive viewing conditions where isolated objects, pairs of objects, or triplets of objects are presented randomly at a rate of 5 stimulus conditions per second. This presentation paradigm was used to isolate the feed-forward component of the IT response to visual stimuli (i.e. little time for top-down attentional shifts) and to speed data collection. In addition, IT neurons show robust object selectivity at this presentation rate and it is consistent with the presentation rate produced spontaneously by free viewing monkeys. These data have allowed us to systematically examine and characterize the effect of placing distractor objects in each IT neuron's RF while a 'preferred' object is also present. In contrast to the predictions (in our proposal) suggested from the original version of the standard model, this is true even for distractor objects that are very dissimilar from the 'preferred' object and produce no detectable response when presented in isolation. As expected from the model and from previous

experimental data, most IT neurons are not idealized object detectors in that they do not respond equally to preferred objects in spite of the presence of non-preferred objects. Our results show that, across a wide range of stimulus conditions, IT neuronal responses to multiple objects – averaged across neurons -- are very well predicted by the average of their responses to the constituent objects. The observations about the average response are consistent with various phenomenological models (without a computational motivation) in which the output of each IT neuron is normalized by, e.g., the total synaptic drive into IT and/or the total spiking activity out of IT. However, the observed averaging effect poses strong additional constraints on computational models of recognition and in particular on the new version of the standard model (described earlier) which attempts to explain properties of visual recognition together with properties of neurons in intermediate areas (V1, V2, V4, IT). In particular, the experimental data on the averaging effect are likely to provide constraints on the form of the normalization operation.

Aim A2: Because we sought to first understand the 'baseline' clutter tolerance properties of IT neurons, we have not given our monkeys extensive experience in the same recognition tasks in the presence of distractor objects or other clutter. However, to lay the groundwork for understanding the effect of experience on IT clutter tolerance, we have: 1) characterized 'baseline' IT responses in clutter (animals not trained in clutter, described above); and 2) made a significant extension of the model to account for the effect of visual experience in the tuning of neurons from area V4 to IT.

In terms of the model, we have developed a new, more detailed version involving learning and similar but more specific mechanisms for max and tuning**.** Instead of the hardwired units of the original model, we assume that simple (S2) units and complex (C2) units (corresponding to V4 in the model) become tuned to patches of C1 units (corresponding to complex cells in V1/V2) activity that repeat across different images of the same objects. We have simulated a simplified version of Foldiak's trace rule to generate S2 and C2 cells that become tuned to complex features of images. After presentation of many natural images, the units become tuned to complex features ─ for instance of face-components ─ if a sequence of face images (in the presence of background) is presented (in general objects are not at the same position and scale). Learning is task-independent and simply relies on temporal continuity (*e.g.* the same object being present during a temporal sequence of images). The same process is iterated in PIT (S3 and C3 cells) where now the neurons become tuned to patches of activities in V4. We also assume, consistently with available data, that there are direct projections from V2 (roughly corresponding to the S1 and C1 cells of the model) to PIT generating S2b units with more selectivity (they are tuned to larger patches with a larger number of subunits than the S2 units in V4). Interestingly the model with its associated learning rules creates a redundant dictionary of features in PIT with different degrees of complexity/selectivity and invariance. For instance, PIT neurons receiving direct projections from V2 are tuned to complex features learned from experience and consisting of configurations of several subunits of the V1 type (each one with a limited range of scale and position invariance, similarly to complex cells in V1). Projections from V4 to PIT support simpler feature cells with a larger degree of invariances.

Aim B1: We examined a population of 144 PFC neurons and 151 ITC neurons while monkeys alternated between categorizing our morph stimuli into "cats" versus "dogs" and matching specific, individual, category members. As in previous studies, we found that the activity of some PFC neurons encoded category, others the specific individuals. The majority of PFC neurons showed similar activity across the two tasks. This suggests common substrates for each task or perhaps adoption of a "hybrid" strategy in which the monkeys simultaneously encode the category membership as well as the identity of individuals (Freedman, Serre, Riesenhuber, Poggio, Miller). We also tested for task specificity by recording from PFC and ITC neurons while monkeys performed our categorization task versus passive viewing of the same stimuli. Task versus passive viewing had a large effect on PFC, but not ITC activity.

Aim B2: Monkeys are currently being prepared for this aim (Roy, Miller). This task requires that monkeys be overtrained on the categorization task in order to produce effects of categorical

perception. So, the monkeys used for Aim B1 will be used for this aim once the experiments for Aim B1 are completed.

3. Plans

Aim 1: We have prepared a manuscript describing the effect of multiple objects on the responses of IT neurons in both arbitrary and parameterized shape spaces (describe above, to be submitted in the next two weeks). Although these data shed new light on the rapid, feed-forward representation of multiple stimuli (clutter) in IT, and some of the results agree with the model some do not agree with the predictions in the original proposal. Specifically, almost all IT neurons studied thus far show response suppression when a second (distractor) object is presented along with a preferred object, no matter how dissimilar the distractor object is from the preferred object, while versions of the model predicted recovery from suppression for dissimilar distractor objects. There are several non-exclusive possibilities that could explain this discrepancy between the data collected thus far and the model: 1) *physiology*: perhaps IT neurons show different clutter tolerance properties when the animal is actively performing a recognition task (relative to the passive viewing conditions we have used so far), 2) *physiology*: perhaps we have not yet recorded from neurons using their true 'preferred' objects (the conditions in which the model most robustly predicts no interference from other, dissimilar objects), 3) *physiology*: perhaps, by focusing on neurons with sharp selectivity (tuning within an object class), we have missed clutter tolerance effects seen in neurons with broader selectivity, 4) *computational*. The standard MAX model is incorrect in some assumptions and/or missing important components that exist in the brain.

We are re-examining and re-verifying the conditions and the assumptions under which the rebound effect will be measurable. We are currently pursuing each of these possibilities. Specifically, we are: 1) testing IT neurons under the same stimulus conditions but with the animal actively performing the target detection task; 2) continuing the search for neurons that are relevant for the recognition task the monkey was trained on (e.g., by showing selectivity for one of the target objects the monkey was trained to discriminate from distractors), as these neurons are most suitable to test the model predictions; 3) investigating the clutter tolerance properties of IT neurons with broader selectivity (e.g., responding to well to *all* faces but not to cars); 4) measuring the monkeys' psychometric curves for recognition in clutter to establish whether the monkeys are in fact able to perform the task.

We expect these recording to be largely complete in the first animal in the next several months. Even more critical than the U curve effect as a constraint on the class of plausible models is the average response of most neurons in IT. We know that both the old and the new version of the model show an average effect in IT for toy stimuli as an effect of both the tuning and the max-like operations; we also know that the original version of the model gives only an approximate average for the real stimuli (so the agreement with the data is not very good), but we do not know yet how well and under which conditions the new version of the model with the normalization-based tuning (see aim N1 above) may be consistent with the IT clutter "averaging" effect described above. We expect to explore the new version of the standard model and constrain it appropriately exploiting these data in a few weeks.

Aim A.2: We still plan to train each monkey to detect the target shapes in the presence of flanking distractor shapes (clutter). However, because of the unexpected observations described in Aim A1 (above), we have focused our efforts on understanding IT response properties in clutter without extensive training (Aim A1).

CalTech Progress Report Summary - 2005

1. Specific Aims

The Caltech project is organized around the central theme of attentional aspects of object recognition, using visual psychophysics, single cell electrophysiology in the human medial temporal lobe, and computational and biophysical modeling. The research is organized into three aims: (1) Psychophysics of attention and recognition in natural scenes parallels electrophysiological work, using both familiar and unfamiliar stimulus categories, to better understand how attention acts at the neuronal level. This will determine the limits of the current feed-forward recognition and saliency models. (2) Integrate our saliency model with the feed-forward recognition system central to our Conte Center to implement attentional modulation of object recognition. (3) Ongoing work with Poggio's group seeks to investigate single neuron and network models of how the MAX operation could be carried out in cortex, aiming to account for the results obtained by the Ferster group.

2. Studies and Results

Face Identification without Engaging Spatial Attention: The processing of naturalistic stimuli has come under a fair amount of attention (Li et al, 2002; Rousselet *et al*., 2002; Braun, 2003; Kayser, 2004). Li and colleagues showed that the visual system can categorize natural scenes more efficiently than artificial geometric shapes, concluding that the attentional demands of a task are not determined by the complexity of the stimuli used, but by their nature – natural scenes versus artificial stimuli. We reported that this type of pre-attentive processing of natural stimuli extends to discriminating the gender of a face – a task that involves discriminating between stimuli which share the same features and only differ subtly in the spatial arrangement of these features (Reddy *et al.*, 2004). These results are in contrast to the finding that subjects cannot distinguish between rotated letters (e.g. 'T' versus 'L'), or a red-green bisected disk from its mirror image, under similar conditions. In the present study, we investigate whether even finer discriminations, at the level of the individual, can be performed in the near-absence of attention.

Using the dual-task paradigm, in which subjects perform face-identification along with a task that is known to remove attention, we test whether subjects' performance on face-identification suffers when attention is removed. Subjects were required to perform face identification on a set of faces of current celebrities, as well as faces of unknown individuals (whom the subjects were exposed to for just 30s prior to the start of the experiment). In the near-absence of attention, performance on this task is not significantly impaired for all but two of our subjects, on the set of famous faces as well as on faces that subjects are not familiar with. This is surprising considering that subjects make their decision about the identity of the face based on viewing each face for less than 200ms on each trial. While performing face identification on a set of inverted faces, the subjects' performance was significantly impaired in the near-absence of attention. These results thus suggest that visual system is able to make complex judgments of natural stimuli, even when attention is not fully available. Neuronal populations which show a high degree of specificity for famous individuals or buildings (Quian-Quiroga *et al*., 2005), could form the basis for the high performance we observe on the celebrity identification task in the absence of attention.

When does visual attention modulate hemodynamic activity in cortex? Many studies have reported that hemodynamic activity in visual cortex is reduced in the absence of selective attention. These reports are at odds with psychophysical data showing that observers are able to efficiently categorize natural stimuli outside the focus of attention. To reconcile these two lines of evidence, we study the effects of attentional modulation on face-selective responses in the fusiform gyrus using fMRI. Different from previous fMRI studies in which an "attended" condition (where subjects make a behavioral report on targets) is compared to an "unattended" condition (where the targets are behaviorally irrelevant), we included a third condition in which the targets were outside the spatial focus of selective attention yet remained behaviorally relevant, enabling us dissociate behavioral relevance and attentional modulation. Whether or not subject had to spatially attend to the faces in order to discriminate male from female faces

made no difference to the amplitude or time-course of BOLD activity in the fusiform face area (FFA) provided that the faces had to be discriminated. We observed a decrease in BOLD activity in the FFA when faces were behaviorally irrelevant. The modulation of the hemodynamic response as a function of the subject's behavior is region specific, as it does not extend to the parahippocampal place area.

Neural Correlates of Change Blindness in the Human Medial Temporal Lobe: Observers are often unaware of changes made to the visual environment when attention is not focused at the location of the change (*change blindness*). Its correlates at the single cell level remain unclear. We recorded from the medial temporal lobe (MTL) of patients with pharmacologically intractable epilepsy, implanted with depth electrodes and microwires, to localize the focus of seizure onsets. Subjects were presented with one set of 4 simultaneously presented images twice, each time for 1s, with a brief blank interval of 1.5s between the 2 presentations. On half the trials, a change occurred at one of the four locations, and subjects had to report whether they detected the change or not. In separate "screening" sessions, specific images that cells were visually responsive to ("preferred stimuli") were determined. In collaboration with Dr. Itzhak Fried at UCAL, we investigated neuronal responses when the set of preferred stimuli were used as changing elements. We recorded from about 700 cells in 9 patients of which 29 were visually responsive under this paradigm in Dr. Fried's lab. These were located in the amygdala, hippocampus, entorhinal cortex and parahippocampal gyrus. Similar to the finding of the MIT report in their recordings in monkey IT, the majority of our cells in the human MTL respond much less to a preferred stimulus (here natural scenes) when in the presence of three distracting images compared to when the image is presented by itself (the firing rate is reduced by ca 80%). Over these cells, the preferred stimuli elicited significantly higher firing rates on correct trials (e.g. change detection) compared to incorrect (e.g. change blindness) trials. For each cell, we were able to predict on a trial-by-trial basis (using a ROC analysis) whether or not a change occurred 67% of the time on average. This prediction was significantly higher than chance on correct trials, but on incorrect trials prediction was at chance. On a trial-by-trial basis, we are also able to predict the behavioral decision of the subject above chance (59%; choice probability). Thus, the firing rates of certain MTL cells might constitute a neural correlate of change detection and change blindness.

Task Switching with Top-down Cues: By focusing their visual attention on a given task (e.g. detecting an animal in an image), humans can increase their efficiency in this task compared to the naïve condition. How much time does it take to load the necessary instructions to bias the attention system effectively? We investigated this question in a task switching paradigm. In this paradigm, the subjects are presented with a grayscale image of a natural scene with a colored frame around the image. The image contains an animal, or a vehicle, or neither (distracter). The frame around the image is orange, blue, or purple. After the task SOA (typically 120 ms) the image and the colored frame were masked. Subjects were trained on four different detection tasks in a block design – "animal", "vehicle", "orange", and "blue". The task SOA and the saturation of the isoluminant colored frames were adjusted such that the subjects achieved between 85% and 95% performance during training. During testing, we introduced a task switching condition in addition to the blocks with individual tasks. For switching blocks, subjects were instructed to solve two out of the four possible tasks during a particular block. In each trial, the subjects had to perform one of the two tasks, which was cued by a brief symbolic cue (a simple geometric shape at fixation). We varied the time between cue onset and stimulus onset (cue SOA) between 800-0 ms and introduced two control conditions in which the cue only appears when the stimulus is replaced by the mask, or 300 ms later when the mask disappears. The subjects were asked to respond only to successful detection of the target by briefly releasing a mouse button. We measured the subjects' performance in the task as well as their reaction times. With these measures, we could compute the switching cost as the difference in performance between task-repeat and task-switch trials in task-switch blocks, and the mixing cost as the difference in performance between task-repeat trials in task-switch blocks and trials in single-task blocks.

We found significant mixing cost of around 100-150 ms in reaction time and around 5% in number of correct trials in all cases. Surprisingly, the tasks could be performed equally well whether the cue was presented 800 ms before the stimulus onset, or only after mask onset. This leads to the conclusion that top-down attention is not a major factor in the performance in these experiments. Rather, subjects appear to be able to hold attributes of the stimulus in memory and make their decision about the correct response after they perceive the cue.

Modeling Feature Sharing between Object Detection and Top-down Attention: Visual search and other attentionally demanding processes are often guided from the top down when a specific task is given (e.g. Wolfe *et al.*, 2004). In the simplified stimuli commonly used in visual search experiments, e.g. red and horizontal bars, the selection of potential features that might be biased for is obvious (by design). In a natural setting with real-world objects, the selection of these features is not obvious, and there is some debate which features can be used for top-down guidance, and how a specific task maps to them (Wolfe and Horowitz, 2004). Learning to detect objects provides the visual system with an effective set of features suitable for the detection task, and with a mapping from these features to an abstract representation of the object.

Together with Poggio's group, we developed a model in which V4-type S2 features are shared between object detection and top-down attention. As the model familiarizes itself with objects, i.e. it learns to detect them, it acquires a representation for features to solve the detection task. We propose that by cortical feedback connections, top-down processes can re-use these same features to bias attention to locations with higher probability of containing the target object. Our implementation of the model outperforms pure bottom-up attention. The performance of our model, which uses only grayscale information, is comparable to a top-down bias for skin hue.

Deployment of Feature-Based Top-Down Attention in Visual Search: Artificial search arrays are used to investigate which features and combinations thereof attract top-down attention (pop-out). If the combination of features of the target does not pop-out, the reaction time (RT) for finding the target increases proportionally as a function of the number of distractors. This is a consequence of the requirement of some sort of serial scanning of some of the distractors. Multiple principal serial search strategies could be utilized: random search, serial search of all distractors which share one or multiple features with the target or a combination thereof. If the search strategy for a known target is to serially search all the items that share a certain feature dimension with the target, this would allow us to investigate which features are used for top-down deployment of attention in visual search.

We recorded eye movements while subjects searched for a known target in search arrays composed of colored oriented bars. At the beginning of each trial, the target alone is shown at the center of the screen, followed by a search array composed of 49 items, one of which is always the target. Subjects (n=5) took on average 8 fixations to find the target. We constructed a conservative computational model with perfect memory for all previous saccades to generate realistic, but random, saccades. The computational model requires on average 25 fixations to find the target, while the subjects require 8 on average. In most trials, subjects serially search the distractors which share one of the features (e.g., color or orientation) with the target. Surprisingly, color turns out to be the dominant feature: even if the two colors are very close (hard to distinguish) and the two orientations used are as different as possible (0 and 90 degrees), color is used as the feature for top-down attentional deployment. We repeated the same search experiment for a large number of trials on the same subjects over multiple days. The search strategy used for a given search array is highly consistent, both for the same subject over a long period of time as well as across subjects. We thus conclude that the feature sets used for top-down deployment are highly stereotypic.

3. Plans

We will continue our research as outlined in the original proposal. As a new aim, in the coming year, our group, in conjunction with Poggio's, are considering the type of operations that can be performed in feed-forward, hierarchical networks, such as the standard model, with and without attention and eye movements. We make several assumptions. (1) The only role of eye movements is to bring a particular part of the retinal input under the high resolution at the fovea. We further assume that the network has been trained for a particular task. This mimics the role of expectation. It is known from the work of Mack and Rock (*in attentional blindness*) that an unexpected visual stimulus may be perceptually invisible even though the observer is directly looking at it.

We define a discrimination task as immediate when it can be carried out with a single eye fixation and in the (near) absence of spatial, focal attention. Under these conditions, we conjecture (1) that such a

simple discrimination (including that between natural scenes containing animals and non-animals) can be done in a feed-forward architecture; (2) that when a task can only be performed with spatial attention (such as telling a red-green from a green-red disk in the periphery) or using multiple fixations, feedback is essential. This may requires additional processing time. (3) Discrimination tasks that fall within the equivalence class defined by some particular invariance require selective, focal attention.

Georgetown Progress Report Summary - 2005

1. Specific Aims

The Georgetown subcontract was originally part of the MIT project, the subcontract arising from Dr. Riesenhuber's move to Georgetown University to start a lab there. Thus, activities at Georgetown are directly related to the aims of the original MIT project, *i.e.,* its Aims A (investigating the neural mechanisms underlying object recognition in clutter) and B (to determine if there is a common neural substrate for different recognition tasks). In parallel, we have amplified the original aims by exploring extensions of the computational model at the core of our center to quantitatively model not just physiological data but also human object recognition performance and brain activity as measured by fMRI. We chose to develop this approach with the object class of human faces. This object class is particularly interesting, not just because of its great importance for cognition, but also because current theories of human object recognition ascribe a special status for faces, claiming that the neural mechanisms underlying face recognition differ from those for other, generic objects, *e.g.,* by using "configural" processing. Our modeling results (submitted for publication) indicate that the same model that drives the physiological experiments can also quantitatively account for human face processing performance and brain activation as measured by fMRI. We then used our computational model of human face processing to generate novel, quantitative predictions for psychophysics and fMRI based on simulated "face cell" activations, with very encouraging preliminary results (see below). This is quite exciting as it significantly broadens the scope of the model and opens the door to use the computational model as a tool to go beyond current "black box" models of neural disorders involving object recognition deficits (such as autism, dyslexia or schizophrenia), with the goal of mechanistically linking behavioral differences to differences at the level of neural processing, thus identifying potential targets for therapeutic intervention. The preliminary results have been so encouraging that we have started a collaboration with a neurologist at Georgetown University Medical Center with the aim of applying this approach to study the neural mechanisms underlying face processing deficits in autism.

2. Studies and Results

Aim A.1: We have continued testing the model predictions regarding recognition in clutter using human psychophysical experiments. In our earlier studies we had trained subjects on novel objects and then tested subjects' recognition performance for these objects presented together with other, distractor objects whose similarity to the target object was parametrically controlled. Key advantages of this approach are its good control over stimulus set and prior experience with the stimuli, and we collaborated with DiCarlo's group to develop a similar paradigm for the physiology project based on the human pilot experiment. A drawback of the "novel object" approach, however, in particular for human studies is that it requires extensive training. In the past year, we thus focused our efforts on an object class for which every human is an "expert" (and thus does not require training): faces. Based on published physiological data on monkey face cell tuning specificity, brain imaging data on FFA (fusiform face area, a brain area crucial to human face perception) selectivity from fMRI, and human behavioral data on face discrimination performance, we have developed a computational model of face neurons in the FFA. The simulations show that the data on face processing can be well accounted for in our standard modeling framework as a result of extensive experience with faces, without having to postulate additional "face-specific" mechanisms, in line with our earlier psychophysical results which were published in the reporting period (Riesenhuber *et al.* 2004). The modeling results have now been submitted for publication (Rosen & Riesenhuber). We are currently testing novel model predictions regarding a quantitative link between face neuron responses, BOLD contrast in fMRI, and behavior. This approach is illustrated in Figure 1 (Jiang & Riesenhuber). In addition, we are conducting psychophysical experiments on face perception in clutter (Jarudi, Jiang, Riesenhuber). Preliminary results (based on 8 subjects) indicate that target detection performance as a function of similarity of target and simultaneously presented clutter object might follow a U-shape, as predicted in the proposal.

Aim B.1: We have collaborated with Earl Miller and Jefferson Roy to develop an experimental paradigm and stimuli to train and test monkeys on a category switching task (see MIT progress report) which will allow us to test model predictions regarding the neural mechanisms underlying different recognition tasks, in particular the split into a generic shape-based representation for the target object class providing input to task-specific circuits. Neural recordings have been completed and data are currently being analyzed. In parallel, we have developed a human subject version of the original categorization task, using morphed cars instead of the cat/dog morphs employed in the monkey study, to avoid confounds with subjects' preexisting categories for the animal stimuli. Using a morph space spanned by four prototypes (the same size currently being used in the monkey studies), we have now trained 12 subjects (for up to nine hours each) on a categorization task, using an optimized training procedure. After training, subjects were able to categorize the morphed stimuli at high accuracy. We are currently investigating categorical perception effects to inform the design of the monkey paradigm for Aim B.2 (Jiang & Riesenhuber).

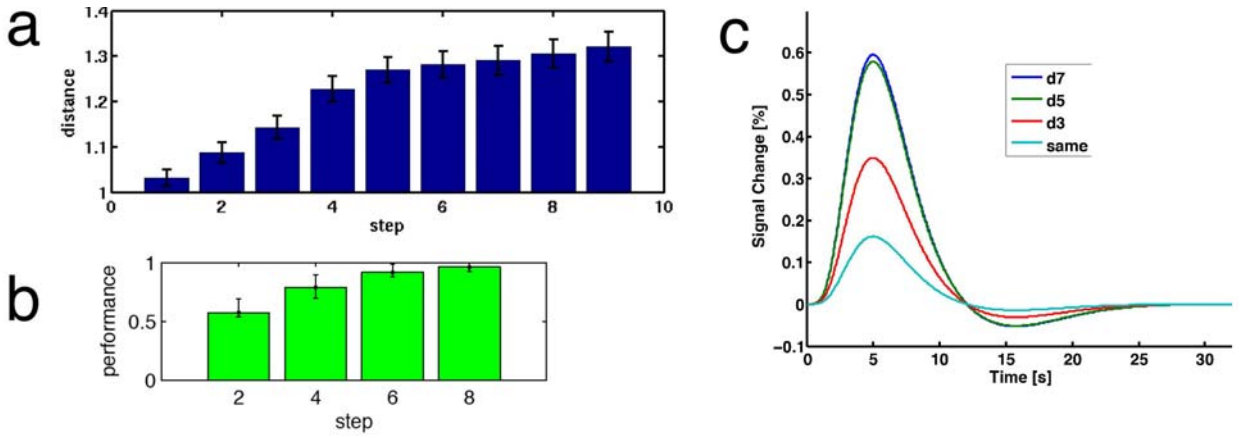

Figure 1: Illustration of the integrative approach of predicting human behavioral performance and fMRI BOLD contrast based on our model of face neurons (FN). ***(a)*** *Average distance (with standard error) between model face unit (FU) activity patterns for faces images of varying similarity. Face pairs were created using a photo-quality morphing systems (Vetter & Blanz, 1999). x-axis gives intra-pair dissimilarity (in equidistant "steps" in morph space, with 10 steps (=d10) corresponding to the distance between two different face prototypes and intermediate steps corresponding to faces that are morphs of the two prototypes), y-axis shows difference between activity patterns over 182 FU not including units tuned to any of the morphed faces. The model predicts a direct link between FN tuning specificity and discrimination performance. If face discrimination is based on the comparisons of FN activation patterns, performance should increase with dissimilarity between target (T) and distractor (D) faces, as the corresponding activation patterns get increasingly dissimilar. Crucially, due to the tight tuning of FNs, for some T-D dissimilarity, both will activate disjoint subpopulations of FNs, and performance should asymptote, as further increasing the T-D dissimilarity will not increase the dissimilarity of FN activation patterns. Likewise, in an fMRI rapid adaptation paradigm (fMRI-RA), adaptation of FFA FN stimulated with pairs of faces of increasing dissimilarity should decrease, and asymptote when the faces activate different subpopulations of FN. Using the results in (a), we quantitatively predicted the T-D dissimilarity for which BOLD-contrast and behavior were expected to asymptote. Crucially, activity pattern distance appears to saturate at around d5, with the distances at d5 and d7 (or d6 and d8) not statistically significantly different, $p>0.26$ ($p>0.14$), whereas d3 and d5 (d4 and d6), although of equal physical dissimilarity, were significantly different, $p<0.0001$ ($p<0.008$).* ***(b)*** *shows experimental 2AFC face discrimination results (9 subjects). Subject performance asymptotes around morph step 6, in close agreement with the prediction in (a).* ***(c)*** *Time course of BOLD-contrast (% signal change relative to a fixation baseline) in one pilot subject's FFA for an fMRI-RA experiment which measures BOLD contrast response to rapidly presented pairs of faces of varying dissimilarity (3,5, and 7 morph steps apart). The different curves show responses to image pairs of different similarity. Very importantly, the BOLD contrast signal saturates around d5, in quantitative agreement with the predictions in (a).*

3. Plans

Aim A: We will continue the psychophysical testing of the model predictions for recognition in clutter, in particular the U-shape prediction for recognition performance. A second focus will be to investigate learning effects: If subjects are trained on a novel object class, how robust is their recognition performance to clutter? Can it be improved through training? The model predicts that improving recognition performance in clutter requires learning at intermediate levels of the visual system (e.g., V4/PIT, see the results by Serre and Poggio in the MIT project). We also intend to explore this question in fMRI (funded separately), with significant synergies for the Conte project, in particular the physiology studies of DiCarlo, and the psychophysics of Koch.

Aim B: We will continue the collaboration with Miller's group, and also continue the human studies on categorization and categorical perception. We also intend to perform fMRI studies (funded separately) of the human subjects trained on the categorization task and relate the findings to the monkey results.

Northwestern Progress Report Summary - 2005

1. Specific Aims

We continue to study the summation of signals within the receptive fields of complex cells in area 17 of cat visual cortex, and the mechanisms underlying the summation properties we observe. As discussed in the last progress report, our early data, now published in the Journal of Neurophysiology (2004 92:2704-13) were somewhat at odds with one of the classical papers on complex cells by Movshon *et al.* (J Physiol. 1978 283:79-99). These authors showed that spatial interactions between pairs of pairs flashed in the receptive fields of complex cells showed a substructure that resembled the receptive fields of simple cells. That is, for flashed bars of the same polarity (bright/bright or dark/dark), when they were close together, they facilitated one another and when they were far apart they antagonized one another. The opposite occurred when the bars were of opposite polarity to one another (bright/dark). We, on the other hand, had found that the interaction between bars in the pair was independent of polarity or separation, and that the interaction was MAX-like in that the response to the pair was similar to largest of the responses to the individual bars. In our experiments, however, we had not tested a complete set of bar pairs at all possible locations, but had tested a few locations in each cell and averaged the results together.

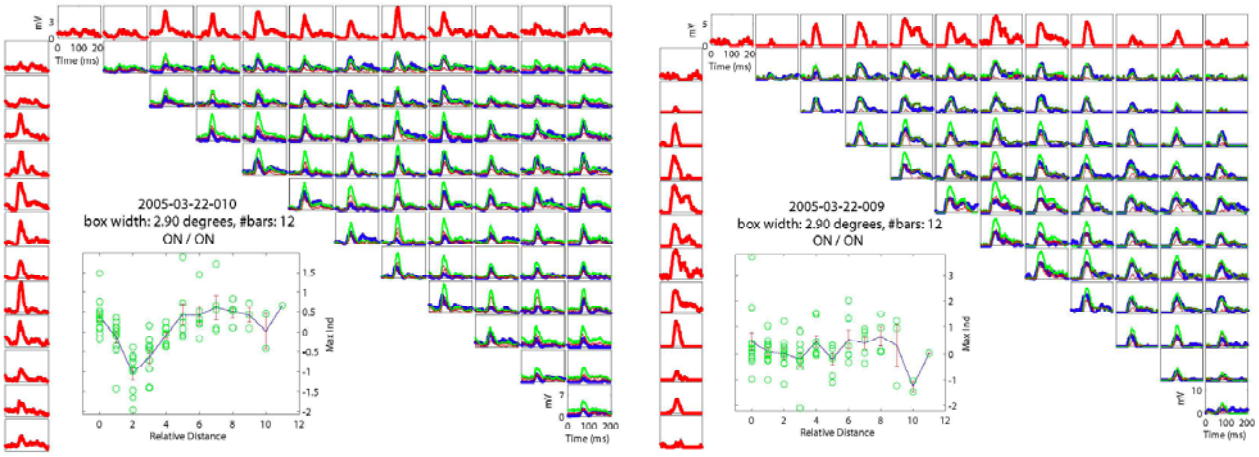

Figure 1. Intracellularly recorded responses of 2 complex cells in area V1 of the cat visual cortex. Top row and left column: responses to a briefly flashed, optimally oriented bar at 12 different positions within the receptive field. The rest of matrix shows the responses to simultaneously flashed pairs of bars (blue). Green traces show the sum of the responses to the two bars of the pair when presented individually. The inset shows the MAX index as a function of the relative distance between the bars in a pair for all pairs.

2. Studies and Results

In order to explore the apparent discrepancy between our data and the Movshon *et al.* data, we have now recorded from a number of complex cells and flashed pairs of bars at all possible combinations of receptive field positions. Two example cells are shown in Figure 1. The response (intracellularly recorded membrane potential) to single bars flashed at 12 positions across the receptive field are shown along the top row (red) and repeated along the left side. The matrix of traces shows the responses to each possible pair of bars. The red traces in each cell of the matrix are the individual responses; the green trace is the arithmetic sum of the individual responses; the blue trace is the actual response to the pair. The cell on the right shows the MAX-like behavior that we described previously. No matter what the position of the bars in the pair, the summation is MAX-like in that the response to a pair is generally

comparable to the larger of the two individual responses. When the MAX index is plotted as a function of bar separation (inset), the plot is relatively flat and averages near 0.

The cell on the left behaves more as Movshon *et al.* described complex cells. The MAX index in this case is clearly dependent on the separation distance between the bars. The negative indices at small separations indicate suppression such that the response to the pair is smaller than either of the individual responses. And the positive indices at larger separations indicate facilitation or summation between the stimuli such that the responses are larger than either of the two individual responses (though somewhat smaller than their sum).

The two cell types that are represented in Figure 1 also differ in their spatial frequency selectivity, as predicted by their spatial interaction. In Figure 2, the spatial frequency selectivity of the left- and right-hand cells in Figure 1 are the 3rd and 2nd plots from the top. The cell with uniform MAX-like behavior has broader spatial frequency tuning and a lower peak spatial frequency, as would be predicted by a Fourier transform of the spatial interaction profile.

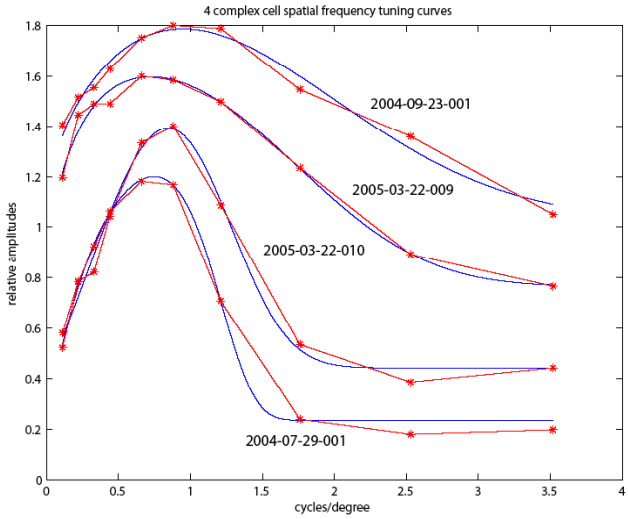

Figure 2. Spatial frequency tuning curves from 4 cells in area V1 of the cat visual cortex. Data are normalized and shifted vertically for clarity. Measurements are of mean depolarization of membrane potential evoked by drifting gratings of different spatial frequencies. Red curves, data; blue curves, fits. The second and third curves from the top are for the right-hand and left-hand cells in Figure 1.

3. Significance

From these results, we can state with some confidence that there really is a population of cells in the visual cortex that specifically perform a MAX-like computation on the visual image. These cells are distinct from the cells that are more classically defined as complex. And, the MAX-like cells may have distinct receptive field properties from the classically-defined complex cells.

4. Plans

It seems then, that Movshon *et al*. reported on only a subset of complex cells in area 17 and that there is an additional population with very different properties than the ones they described. We will continue to explore these cell types and their differences. We are interested in whether these two different cell types lie at the ends of a continuum or constitute two distinct populations, where (what layers) they are located in the cortex, whether they differ in receptive field properties such as size, orientation specificity, direction selectivity, length summation, and whether they differ in synaptic connectivity. To test the latter, we will

place a stimulating electrode in the lateral geniculate nucleus and measure the latency of evoked PSPs. Shorter latencies (< 2.3 ms) are indicative of direct, monosynaptic connections. Longer latencies (> 3.0 ms) are indicative of indirect, or polysynaptic connections via other cortical neurons. We will also be able to identify, though the antidromic responses of the cells, whether they project to subcortical structures such as the lateral geniculate or superior colliculus.

In addition we will explore the cellular mechanisms underlying the different types of summation in these cells. By recording the visually evoked changes in membrane potential while polarizing the cell with different levels of injected current, we can estimate the excitatory and inhibitory synaptic conductances underlying the potentials. These conductances will likely differ qualitatively in the two types of cells. Our colleague Ilan Lampl at the Wiezmann Institute, for example, suggests that MAX-like behavior might be produced by nonlinear summation of inhibitory inputs combined with linear summation of excitatory inputs. Other models of the mechanisms underlying the MAX-like behavior of complex cells are being developed by members of Dr. Poggio's and Dr. Koch's groups. Both the intrinsic properties of cortical neurons (synaptic and voltage gated currents) and the properties of the local circuit are being considered. Where possible, our experiments will be designed to test these models specifically.