Standing, left to right:
A. Hurlbert, H.H. Bülthoff, J. Altman, C. von der Malsburg, N. Logothetis, W. Singer, J.H.R. Maunsell, H.D. Mumford

Seated, left to right:
S. Ullman, D.A. Glaser, K.A.C. Martin, K. Hepp

15

Group Report: Vision—A Case Study in Brain Function

or

"It's a Brain-damaged Group that Can't Recognize a Horace from a Wolf"
—L. Segel, 1991

A.C. HURLBERT, Rapporteur

H.H. BÜLTHOFF, D.A. GLASER, K. HEPP,
N. LOGOTHETIS, K.A.C. MARTIN,
J.H.R. MAUNSELL, D.B. MUMFORD,
W. SINGER, O. SPORNS, S. ULLMAN,
C. VON DER MALSBURG

In the act of seeing, we segment the world into objects. Inherent in this are two processes, inextricably linked: demarcation and grouping together of parts or features that consitute an object, and recognition of the object. Here we will discuss the state of our understanding of how the human visual system performs these formidable tasks. What little understanding we have of these processes exists on several levels, in the form of experimental data and theoretical models. The data come largely from studies of the electrophysiological behavior and anatomical connections of individual cells in the visual systems primarily of monkey and cat, and from the performance of normal and brain-lesioned humans on visual tasks. Some models have been formulated explicitly to be implemented on a computer connected to a camera and thereby to equip a machine with vision; others attempt to collate and explain biological data. The distinctions between the two classes are not always precise: some computational models also serve as a test for assumptions about the constraints under which the human visual system may work and some models designed to explain biological data may be tested by computer simulations. The purpose of this workshop was to bring together theoreticians and experimentalists who were willing to blur the boundaries between their approaches in order to sharpen our understanding of vision.

Exploring Brain Functions: Models in Neuroscience
Edited by T.A. Poggio and D.A. Glaser © 1993 John Wiley & Sons Ltd

OBJECT RECOGNITION

The Viewpoint of Machines

Artificial visual systems are capable of impressive recognition tasks, despite the fact that we understand very little of how the human visual system recognizes objects. Most models of object recognition make the obvious assumption that the visual system (of man, monkey or machine) stores representations of objects and that recognition occurs when a representation is matched to an incoming stimulus. This assumption raises two questions: what is the nature of the stored representation?; and how is matching achieved?

In machine vision, object representations take several forms. Objects may be represented pictorially, as the initial image itself, whole or in part, or as a contour map of the initial image. On increasing levels of abstraction, objects may be represented as lists of features extracted from the image. The features may simply be corners, blobs, or line terminators, or they may be areas, moments, or other descriptors designed to be invariant under certain translocations or transformations that the object might undergo. In other models, objects are decomposed into 3D parts such as cylinders, cones, and cubes and represented by descriptions of the structural relations between these parts. In all these representations, there is an obvious prejudice for geometric form. Indeed, most computational models assume that the shape of an object is the critical feature for its recognition. Additional features such as color or texture, or more generally, the material properties of the object, play a small role in object recognition by vision machines. This bias reflects the primacy of shape in the phenomenology of object recognition by humans.

Perhaps the most striking support for this bias comes from studies of agnosic patients who not only fail to recognize unique faces but also cannot recognize certain objects at a nonunique level (Damasio et al. 1990a; Damasio et al. 1990b). Although these are usually natural objects such as animals (a wolf, raccoon, and dachsund might each be identified only as a "kind of dog"), and these patients typically have no trouble identifying man-made objects such as tools or trucks, the underlying explanation for this pattern may not lie in the distinction between natural and man-made. It appears to lie instead in the similarity of shapes: many four-legged animals share similar shapes, but a hammer is geometrically different from a wrench. No prosopagnosiac, however damaged, ever fails to recognize an elephant (whose closest relative, incidentally, is the rock hyrax of Africa, a rabbit-sized mammal). Another way to describe the primacy of shape might be to say that it is most important for categorizing objects on the basic level (dog, horse) whereas material properties and other features are often more important in subordinate-level categorization (my pet Labrador). Color is often more critical in determining individuality than shape, for example, when a red Ford Escort must be found in a crowded parking lot. Recently machine vision algorithms have been developed that exploit the material properties of objects for recognition. In a new method destined for such applications as identifying fruits and vegetables at

supermarket check-out counters, color, not geometric form, is used as an index (Swain and Ballard 1991). The advantage of color lies specifically in the fact that it is not sensitive to viewing angle, object orientation, etc. It can even be made robust under illumination changes when algorithms for color constancy are incorporated. Studies of human vision also suggest that color can serve as a cue to recognition, especially when other cues are sparse or subliminal (Kumar et al. 1992).

In machine vision, the problem of how to match images of new objects to the stored representation depends largely on how complete, specific and robust the representation is. A single rigid template will not match all samples of an object whose images vary substantially with viewpoint or illumination or in more idiosyncratic ways (e.g., a face that can smile or frown). Two different solutions have been proposed to circumvent this obstacle. The first uses a small set of fixed templates (see the next section and Poggio and Girosi, this volume; see also Turk and Pentland 1991 for a related technique, that of averaging pictorial templates). The second relies on a single elastic template that can be deformed to match each new exemplar. Its elasticity may be controlled by varying geometric parameters (Fischler and Elshlager 1973; Yuille et al. 1989) or allowing arbitrary small shape deformations (Grenander et al. 1990). A related approach is to organize the templates into a graph, in which the nodes represent feature points or surface elements each visible from many — but not all — viewpoints and the edges provide geometrical relationships of contiguity. Each view of the object is given by a subset of nodes that groups the presently visible surface elements (Koenderinck 1984). In Reiser's fusion graph model (Reiser 1991), many 2D views of one object are fused into a single template. New 2D views are recognized by elastic matching with the appropriate part of the fusion graph. Despite their advantages over rigid templates, even elastic template models may become unwieldably large when used for complex 3D objects. Most elastic template models also incur large overheads at run time, because the energy minimizations involved in elastic matching are computationally greedy.

Invariant from Any Angle

The majority of computational models are further focused on a specific sub-problem of object recognition: viewpoint invariance. A single 3D object can generate a large number of 2D views, each of which should trigger recognition in a robust system. This is an exemplary problem in computational vision, illustrating both the difficulty of the computational problems underlying object recognition and the sort of implications computational models may have for human vision.

In general, solutions to the problem of viewpoint-invariance can be classified in terms of whether or not they require an explicit 3D representation of the object under scrutiny. A single 3D model of an object can be rotated and projected to generate 2D views for matching, thus economizing on storage space while luxuriating in computational burden. Recently several schemes have been proposed that are relatively inexpensive in terms of both computation and memory. These schemes require only

a small collection of 2D views of each object to generate novel 2D views of the same object for matching. A model proposed by Ullman and Basri (1990) (see also Ullman, this volume) relies on the proof that any orthographically projected 2D view can be obtained by a linear combination of three distinct 2D views (provided features are conserved between views); the generalized radial basis functions (GRBF) network described by Poggio, et. al. (1990) (see Poggio, this volume) performs a nonlinear interpolation between a larger set of perspectively projected 2D views to generate novel views for matching. (The GRBF network is not restricted to representations of geometric form alone. The input vector can consist of features other than spatial locations of contours.) Like elastic template models, which can also achieve viewpoint invariance, schemes that combine 2D views rely on a set of templates to overcome the problem posed by single rigid template. The degree of viewpoint invariance and the level of recognition attained depend on what the model is designed to do. A GRBF network "trained" to recognize a single specific object from any view can readily distinguish it from other objects, even when the two objects are very similar, for example, two bent paper clips (Poggio and Edelman 1990). In this sense it achieves subordinate-level recognition, but not basic-level recognition, since it cannot register the difference between any other paper clip and a coffee cup. A scheme whose templates include object features that might be common to many objects of the same class (e.g., the elastic template scheme of Reiser 1991) in general could distinguish between objects on the basic level, but not on the subordinate level.

Object Representations in Man and Monkey

These models address a specific question on the nature of the stored representation of the geometric form of objects, leaving aside the larger issue of how information from different modules or modalities is integrated in object representation or recognition. Yet nonetheless these schemes have important implications when taken as models of human vision. They underscore the notion that the brain does not need to create and store 3D representations of objects in order to recognize novel 2D views but that instead a small collection of 2D views will suffice. More specifically, any recognition scheme that interpolates between 2D views makes a prediction that can be directly tested against human performance: novel views taken from within the range of training views will be better recognized than views taken from outside the range. This prediction has in fact been confirmed in comparisons between human and GRBF network performance on wire-frame and solid objects (Bülthoff, chapter 11).

The objects used in these experiments, mangled paper clips or globular masses with multiple protrusions, are arguably "nonsensical" (Gerhardstein and Biederman 1991) and not the type that humans would naturally encounter (unless perhaps in certain histological specialties). To avoid artifacts due to past experience and to ensure that there are measurable variations in performance, the objects must necessarily be novel and difficult to recognize. Thus although the result suggests that humans do not use explicit 3D representations to recognize novel views of novel objects, it does not exclude the possibility that 3D models are used for other types of objects or for other purposes. For example, 3D models may play a role in tasks more difficult than fast recognition, involving the detection of mirror symmetry or mental rotation (Shepherd and Metzler 1971; Bülthoff, this volume), although the evidence for this role is still inconclusive (Tarr and Pinker 1989, Edelman and Weinshall 1991).

The swiftness of recognition of familiar objects (at least at the basic level) could be explained by prior exposure to, and fast interpolation between, a multiplicity of natural views. But an alternative explanation is that most natural objects have structurally well-defined parts that permit viewpoint-invariant representations, unlike "nonsense" objects (Biederman 1987). Objects with bilateral symmetry, such as faces, or objects with a dominant axis, such as bones, trees, and many tools, may be represented in different ways and recognized by other strategies. In contrast to results obtained on "nonsense" objects, observers trained on three-quarter views of previously unfamiliar faces recognized them more readily when they were subsequently presented in frontal views (Carey and Etcoff, pers. comm.). These results are particularly interesting in light of a scheme recently proposed by which a novel view of a bilaterally symmetric object may be recognized from a single nonfrontal model view (Poggio and Vetter 1992). This scheme is closely linked to 2D view interpolation models. Thus it is possible, although only a speculation, that if the visual system employs distinct strategies for distinct recognition tasks, these strategies might be special instances of a more general mechanism.

Grandmothers or Old Ladies?

A key question concerning biological object representations is: are they local or distributed?

The extreme version of a local representation is the grandmother cell (Barlow 1972). In caricature, the grandmother cell is one that responds selectively to a particular view of grandmother wearing a particular expression and sporting a particular coiffure. Such specificity of neuronal response is biologically implausible, since it would quickly lead to a combinatorial explosion. But the grandmother cell in this form is only a "straw woman," portrayed as such largely for the benefit of arguments in favor of distributed representations.

Indeed the grandmother cell first proffered by Barlow (1972) was not nearly so exclusive. This more robust cell signals the presence of grandmother, regardless of her expression or pose, and therefore responds to "all views of grandmother's face" (Barlow 1972). It is only one in a collection of cells that together signal the aggregate of features specific to grandmother, but it or a small subset of similar cells always fire when grandmother comes into view.

2D view-interpolation schemes like the GRBF network discussed in the previous section enable such kinder, gentler alternatives to the strict grandmother-cell hypothesis. The grandmother-detector cell might simply register the output of a GRBF-like network trained to recognize Granny from any viewpoint. Whereas the strict grand-

mother cell model would require a specific, sharply tuned neuron for each 2D view of an object, the GRBF network requires only a few neurons, which could be quite coarsely tuned, each responding maximally to a specific view but substantially to similar views as well. The strict grandmother cell hypothesis corresponds to interval coding, whereas the GRBF or "old lady" cell model corresponds to population coding.

Thus, view-interpolation schemes make the specific prediction that objects may be encoded by a small population of coarsely-tuned neurons. Exactly this sort of specificity has been found in "face" cells of monkey inferotemporal cortex: Perrett et. al. (1985) report cells there that are selectively responsive to face view and the direction of eye gaze; Young and Yamane (1991) found that combinations of the coarsely-tuned responses of only 40 cells in inferotemporal cortex were sufficient to encode a set of 27 faces.

Segmenting the Visual Image

The question of how objects are represented bypasses the more crucial question of how the image is initially segmented into parts or features constituting the object to be represented. The problems encountered in machine vision have hammered home the difficulty of segmenting an image without prior information. If a vision machine is told exactly where the face is in a cluttered image, the problem of recognizing that face is in principle far more tractable than when the machine must discover for itself which of the many blobs is a nose. In machine vision, *image segmentation* is traditionally relegated to early vision and starts with decomposing the image into uniform or homogeneous regions and grouping together parts that have, for example, similar color, texture, or motion, but are spatially disconnected. The gap between low level models of image segmentation and higher level models is still mostly unbridged. Some models do address the problem of image segmentation (see, for example, the saliency network described by Ullman, this volume) but those that deal with object recognition usually start with images already segmented into objects.

In human vision, the problem comes under several headings: figure/ground discrimination, grouping, and binding. But whatever the problem is called, the human visual system has solved it. The visual system "knows" which blobs to group together into a face. It can perceive as coherent wholes objects that are partially occluded by other objects or obscured by visual noise. Some of the cues that aid in determining figural unity are: spatial contiguity, conformity of movement or depth, and uniformity of color and texture (see e.g., the Gestalt laws in Wertheimer 1923). For example, in a field of black dots moving randomly on a white screen, a subset of nearby dots that each move in the same direction with the same velocity will emerge as a coherent figure. Segmentation models for machine vision draw heavily on these Gestalt criteria.

A natural question that arises is: does the human visual system exploit higher-level information when it segments images into objects? Or: is figure/ground discrimination driven by top-down processes? Is grouping solely a bottom-up procedure? Here, the

notion of a hierarchical visual system is hard to escape, imbedded in the distinction between top-down and bottom-up. But top-down influences here have a clear definition: they are the processes responsible for changing the perception of a retinal stimulus that is, at bottom, unchanging.

Figure/ground discrimination in flies and infants seems to be (a) bottom-up and (b) primarily dependent on motion. In flies, the physiological mechanism underlying their ability to detect and orient towards an object that differs from its background only by relative motion is now well understood (Reichardt et al. 1983). It is mediated by specific "figure-detection" cells (Egelhaaf, 1990) that receive input only from lower-level motion-sensitive cells. In more developed organisms (both phylogenetically and ontogenetically), there is more evidence for the influence of top-down processes on segmentation. Prior knowledge can affect the perception of an image that on first view is ambiguous or unstructured. For conditioned observers, the Dalmatian dog emerges instantly from its speckled background in an otherwise ambiguous photograph (Gregory 1970). Other reversals of perception can be triggered by prior knowledge or conscious effort: for example, the back/front alternations of depth of the Necker cube (Gregory 1986); or the concave/convex reversals of 3D shape perceived through the kinetic depth effect (Cumming et al. 1991).

Yet other figure/ground discriminations resist the influence of higher-level knowledge. The fragments of a population of letter "B"s are clearly seen as such when a connected black blob obscures their missing parts, but when the occluder is made invisible, the letters are extremely difficult to recognize (Nakayama and Shimojo 1990). A face partially occluded by horizontal stripes is easily recognized when its parts lie in a depth plane behind the stripes, as if seen through a fence, but not when its parts lie in front of the stripes (Nakayama and Shimojo 1990). Grouping by depth, even when assisted by prior knowledge, is not enough to regenerate the whole percept from its parts.

The Gestalt laws in themselves do not explain how the visual system implements them. Spatial contiguity, for example, may be a criterion for grouping line elements into a single continuous contour, but it is only a rule, not an explanation for how line elements are signalled as being spatially contiguous. This is the problem generally called *binding*: what is the physiological mechanism that links together the activity of cells representing image features that are perceptually grouped? Within vision, binding can occur within cortical areas (when, for example, cells encoding the different hues of a multi-colored object must be linked) or between areas (when the color of an object must be linked to its motion). Binding can also occur between modalities, as when a voice is linked with a face to form a more complex representation of a person.

The Binding Problem

Rosenblatt (1961) proposed a hypothetical vision machine to illustrate the "binding" problem. Imagine a machine that views a screen on which can appear a square or

triangle, say, in either the top or bottom part of the screen. Now imagine that the machine has only 4 neurons, two that encode spatial location (one responds to the presence of any visual stimulus in the top part of the screen; the other to any stimulus in the bottom part of the screen) and two that encode geometric form (one specific for triangles, the other for squares). When the triangle appears in the lower part of the screen, an unambiguous pair of neurons is activated. When a triangle appears in the lower part, simultaneously with a square in the upper, all four neurons are activated. Their response is now ambiguous, as the same four neurons would be activated by a triangle in the upper part of the screen appearing together with a square in the lower part. The machine confronts the binding problem: how does it link the response of the cell encoding a specific spatial location with that encoding a specific shape?

At one extreme, the machine could resort to the "grandmother cell," to bind together the responses of two cells in the more selective response of yet another cell. The machine could create by convergent inputs from the existing cells a new cell responsive only to a triangle located in the lower half of the screen, another responsive only to a square in the upper half, and so forth. This would double the number of cells in the machine's brain, using Rosenblatt's design. If carried to its logical extreme in a machine required to recognize as many objects in as many contexts as does a human, this accretion of specialized cells would lead to a combinatorial explosion.

A more parsimonius solution would be to label the activity of each of the original four cells with a temporal code identifying the group to which it belongs. The triangle neuron and the upper field neuron would carry the same code to signify a triangle in the top half of the screen. Temporal synchrony has been proposed as the optimal method for segregating the activities of simultaneously active, distinct cell groups (von der Malsburg 1992; von der Malsburg and Schneider 1986). Cells responding to features that should be linked together would synchronize their firing, while cell groups responding to distinct objects would desynchronize.

Yet the binding problem as defined by Rosenblatt arises from the ambiguity built into the system at its lowest level. Feature-specific cells generalize over position and spatial registry is therefore lost. Whether the architecture of the primate visual system enables it to escape Rosenblatt's binding problem is still an unanswered question, but some physiologists would argue that the problem could never arise in the primate visual system because spatial registration is supplied by the precise topography of the early visual areas. A 4-neuron network with topographically organized square and triangle detectors could perform flawlessly in its limited world without need for an overt binding mechanism. If the machine were required to recognize more objects, it could possibly avert a combinatorial explosion by resorting to "old-lady cells." As discussed above, this alternative would require population coding, in which a given object is represented by the weighted combination of activities of broadly-tuned cells.

Rosenblatt's statement of the binding problem also assumes that the visual system would need to signal more than one object simultaneously. This need could thwart the straightforward use of a population code if spatially intermingled populations are

simultaneously active. Again, some additional means of distinguishing between the populations must come into play.

One such proposed mechanism, selective attention, answers this problem by assuming that the visual system *cannot* signal more than one object simultaneously. This mechanism labels the activity of a group of cells signifying one object by selectively enhancing that cell group and suppressing the irrelevant activities of others. Coactivation is therefore sufficient to signal coherence. The "spotlight of attention" may further tie an object to a specific spatial location by gating the low-level signals that arrive from different regions of the image. Cells in higher visual areas without explicit topography, such as inferotemporal cortex, may continually update information about the spatial location of line segments, for example, merely by being simultaneously active with the appropriate constellation of cells in lower areas. Evidence that such an attentional mechanism exists and can be disrupted lies in the occurrence of "illusory conjunctions," percepts of combinations of features ascribed to a location or object in which they do not actually exist. These illusions occur when the observer's attention is overloaded or diverted (Treisman and Schmidt 1982). By definition, only one object can be spotlighted at a time, so the problem of disentangling simultaneously active distinct cell groups dissolves.

Binding mechanisms that require temporal synchrony might be placed in the same class as selective attention. Both are temporal mechanisms for segregating neuronal responses. But there is a crucial distinction between mere coactivation and specific synchronization of neuronal firing.

Von der Malsburg (1992) first put forward a general theory for sensory segmentation based on this latter form of temporal synchrony ("spike synchronization") and later applied it to the auditory system (von der Malsburg and Schneider 1986). In this network model of a "neural cocktail-party processor," cells stimulated by sounds with similar amplitude modulations synchronize their activities; cells stimulated by dissimilar sounds desynchronize. Inhibitory subsystems prevent the network from falling into global synchrony. Short-term synaptic changes act to stabilize connections between synchronized cells, and decay quickly when activity ceases.

Sporns et. al. (1989; 1991) proposed a similar model for segmentation by the visual system. In this model, sensory space is represented by a set of cells selective for distinct visual features, e.g., edge orientation, color, or motion. The Gestalt criteria for figural unity are represented by two sets of connections: between cells that refer to the same point in visual space irrespective of feature type, and between cells of the same feature type irrespective of position. The signals from active cells are allowed to fluctuate spontaneously. When stimulated by a coherent figure, groups of cells responding to its distinct features synchronize their activities. The activity of cells that respond to background elements are not correlated with cells that respond to figure elements, even when these elements are of the same feature type.

A model of object recognition designed according to general characteristics of cortical organization and implemented on an artificial neural network illustrates how synchronous acitivities may arise in the stabilization of a distributed representation

(Koerner and Boehme, 1991). Interestingly, this model also requires top-down sequential processing to halt the proliferation of inappropriate parallel representations. Bauer and Krey (1990) demonstrate how artificial neural networks constructed from simple threshold units can learn, recognize and sustain periodic sequences of patterns, which are encoded by collective oscillations at distinct frequencies.

The models of binding by temporal synchrony typically are applied to stimuli with intrinsic temporal patterns (e.g., moving figures in vision, sound spectra in hearing). For these temporally dynamic stimuli, temporal synchrony may emerge naturally as a binding mechanism, or even as merely an epiphenomenon. (But it should be noted that in some models, e.g., Sporns et. al. 1989, the synchronous activity that emerges is not locked to the temporal pattern of the stimulus.) Is temporal synchrony also a natural means for other grouping tasks, such as linking together the spatially distributed but stationary features of the face behind the fence (Nakayama and Shimojo 1990)? There may be an evolutionary argument that justifies the more general use of temporal synchrony. If we assume that coexistence in time of different features is the most primitive evidence of an object and that visual motion analysis is the most primitive perceptual ability, then we may argue that temporal synchronization was initially used as the only grouping criterion and mechanism, and later evolved to implement other, more complex, grouping criteria.

A remaining question is whether selective attention or temporal synchrony is in itself sufficient to implement the grouping rules evident in visual perception. In the model of Sporns, et. al. (1991), connections between cells serve as the substrate for the Gestalt grouping criteria. It is only within this structural framework that the intrinsic dynamic variability of neuronal groups acts to synchronize related cells and to desynchronize unrelated cells and that temporal synchrony thereby serves to implement the Gestalt grouping rules. A related question is whether the structural requirements for temporal synchrony as a general-purpose binding mechanism can be met on the scale of the primate visual system.

Another open question is whether the temporal code requires a deciphering mechanism. If so, this mechanism would probably require appropriate efferent connections from the synchronized cell group. Indeed, one of the arguments for the efficacy of temporal synchrony is that temporally synchronous neurons may elicit a larger response from another neuron or group of neurons onto which they commonly converge (see e.g., Sporns et. al. 1992). But postulating an output neuron or group of neurons (which might be, for example, effector neurons in another modality) could reintroduce the "grandmother cell" problem. At least, it might obviate the need for the temporal code. The temporal code might be necessary to ensure activation of an output neuron that is selectively responsive to the phases of its inputs. But perhaps equally plausible is an output neuron selectively responsive to the amplitudes of its roughly cotemporaneous inputs. This latter possibility would not require a temporal code in the strict sense discussed here. These are possibilities that remain to be examined in light of the known functional architecture of the primate visual system.

Experimental Evidence for Binding by Temporal Synchrony

What evidence is there that the visual cortex uses a temporal code to segregate functionally distinct, distributed clusters of neuronal activity? What form might this code take and how might it vary between animals? These are questions which are just beginning to be addressed. There is general agreement that a temporal code must involve temporal synchronization of neuronal responses on some time scale. Whether this requires phase-locked periodic firing or synchronization of aperiodic bursts or merely roughly simultaneous co-activation is still debated. The argument for the physiological efficacy of temporal synchrony is strong: it would seem to (1) increase the probability of firing of neurons that receive convergent input from synchronized cells and (2) increase the probability of inducing changes in synaptic strength, particularly as plastic processes have high activation thresholds (Artola et al. 1990). It can be argued further that periodicity, or more generally, predictability of inter-burst intervals, could aid in establishing synchrony, particularly in the face of often considerable conduction delays in coupling connections. Although periodicity in itself is neither necessary nor sufficient to establish a binding code, the visual system might exploit intrinsic periodic behavior to signal coherent activity.

The recent discovery of stimulus-specific oscillations supports the arguments for periodicity of the temporal code. In anesthetized cat visual cortex, neighboring and distant cells of like specificity (Gray and Singer 1989; Tso et al. 1986; Gray et al. 1989) tend to respond to optimal stimuli with synchronized oscillatory discharges peaking around 40 Hz, most evident in correlograms of local field potentials (LFPs) and multi-unit activities (MUAs). The strongest indication that the observed oscillatory activity in anesthetized cat is functionally significant comes from a recent study by Engel et al. (1991), in which broadly-tuned orientation-specific cells with overlapping receptive fields synchronized their oscillatory discharges when stimulated with a single light bar, but split into groups determined by orientation preference when stimulated simultaneously with two bars of distinct orientations. Within the groups, cell firing was synchronized, whereas between groups it was not.

These and other reports of neuronal response oscillations at similar frequencies in other animals generated enthusiasm specifically for 40 Hz oscillations as the neuronal binding mechanism and even as the substrate of consciousness (Crick and Koch 1990). Yet subsequent investigations have indicated a lesser role for 40Hz oscillations, particularly in monkeys. Newsome and Koch (pers. comm.) analyzed single neuron responses in area MT of alert monkeys performing the task of detecting motion of a group of coherently moving dots against a background of randomly moving dots. The monkey's performance, which improves with the percentage of coherently moving dots, can be predicted from the response curves of single cells. Yet there was no correlation between power at 40 Hz, or in any other narrow band, in autocorrelograms of single cells and the monkey's performance. A small percentage of cells, most of which also tended to burst, did exhibit an increase in power at around 40 Hz, but this

was not necessarily related to the stimulus. Although this result suggests that narrow-band oscillations do not subserve perceptual tasks requiring attention and grouping, it is based on single cell autocorrelograms, in which sampling problems may obscure evidence of oscillations and which could not in any case demonstrate synchrony between cells.

Young et al. (1992) found no evidence for stimulus-related oscillations in autocorrelograms of MUAs and power spectra of LFPs in monkey V1 or MT, using stimuli and anesthesia conditions very similar to those under which oscillations were revealed in the cat. A small proportion of MUA autocorrelograms (2/50) from IT in an alert monkey performing a face discrimination task did show power around 40 Hz, but only one such response was associated with stimulation. Different results are found in nonvisual cortex. In sensorimotor cortex of an alert monkey performing a difficult motor task, Murthy and Fetz (1991) found a higher proportion of transient oscillatory synchronizations in LFPs. Young et al. further suggest that previous estimates of up to 68 % for the percentage of oscillatory responses found in cat visual cortex may have been inflated by over-tolerant acceptance criteria, and that the true proportion may lie between 10 and 35 % (as in Toyoma et al. 1981; Ghose and Freeman 1990). These data illuminate the necessity for standard experimental and statistical procedures to be applied in classifying temporal response patterns.

An unexplained gap still remains between the prevalences of periodic neuronal responses in cat and monkey, which suggests the existence of fundamental differences between cortical structure in the two animals. Clues to the crucial differences lie in a simulation of a cortical neural network (Bush and Douglas 1991) in which the addition of global feedback inhibition was found necessary to synchronize the activity of intrinsically bursting, mutually excitatory cells. Koch and Schuster (1991) found that synchronized bursting in a similar artificial network produced as epiphenomena damped oscillations in cross-correlograms. The structure of cortical connections may be different in cat (Toyoma et al. 1981; T'so et al. 1986) and monkey (Gochin et al. 1991). Reasoning from population dynamics, Young (pers. comm.) concludes that two factors alone may determine where the intrinsic group behavior of cortical neurons falls in the range from periodic to chaotic: the reach of inhibitory connections and the average time delay between inhibition and excitation. In the monkey, synchronized group behavior may lie well beyond the periodic domain.

These observations emphasize the need to shift the focus of the search for temporal patterning away from narrow-band oscillatory activity and towards evidence of more subtle mechanisms of synchronization. Kreiter et al. (1991), for example, report fewer regular oscillations than in cat but nonetheless strong evidence for prominent synchronization in the superior temporal sulcus of alert monkeys. Recent experiments by Vaadia et al. (1991) suggest that correlations between neurons may contribute to the execution of a behavioral task. These authors observed neurons in prefrontal cortex of an alert monkey that synchronized their activity for a short period of time (roughly 200 msec) when the animal initiated a behavioral response. The mean level of discharge of the neurons remained unchanged.

If recurrent bursting underpins temporal synchrony, as some models would suggest, and if the interburst intervals vary widely, autocorrelograms may reveal only broad power spectra of the sort reported in monkey (Young et al. 1992). Cross-correlograms between burst-synchronized neurons may show only single peaks centred on zero time lag. The challenge to experimentalists would then be to demonstrate that this peak does not result merely from common input to the correlated cells, its traditional interpretation.

Binding Levels into Models

Models of the brain and specifically of the visual system have grown dramatically both in number and acceptance since the Dahlem conference in 1977 on "Function and Formation of Neural Systems" where there was only one pure theoretician amongst a large group of experimental neuroscientists. Yet there is still debate on what sort of models are most useful in brain science, on how to evaluate models, and even, on whether models are capable of characterizing a system that might just be a "bag of tricks" (Ramachandran 1985). (Of course, one must remember that "a trick used twice is a method," as a physics professor once admonished students who complained that the solution to an exam problem required memorization of a trick employed in an earlier demonstration (quoted by D. Glaser).)

The primary demand made of a model of the brain is that it be useful; it should either collect and compress data into a comprehensible framework, make specific predictions, or otherwise motivate new experimentation or modeling. Some neuroscientists also believe that the understanding of the brain can only be complete when it encompasses all levels of investigation, from single-cell physiology to neuropsychology to computational analyses, and that therefore the worth of a model can also be judged by the number of levels it spans. Yet even models restricted to single levels may provide powerful insights. In physics, the laws of thermodynamics are on a phenomenological level relative to statistical mechanics. Thermodynamics describes the relationship between the pressure and volume of a gas, for example, without offering an explanation of why a gas exerts pressure. As a phenomenological theory, thermodynamics allowed the behavior of molecules to be exploited before molecules were even known to exist. Of course, the equations of statistical mechanics are themselves phenomenological when formulated in the framework of quantum mechanics; where to draw the line between phenomenological and reductionistic theories is not always clear-cut.

Some would argue, however, that biology is fundamentally different from physics, and whereas a simple, self-consistent model may be accepted by physicists, it would rightly be shunned by biologists. In biology, "Occam's razor can cut the throat" (quoted by L. Segel).

Computational theories of vision arising from the tradition of Artificial Intelligence, for example, have justified themselves with the assumption that the level on which they exist is common to any computing machine, regardless of its components,

including the human brain. Yet the realization underlying more recent computational theories is that perhaps such a level does not exist, because the properties of the components of a machine constrain the very computations it can perform. The storage capacity, temporal resolution, information transmission rate, and sheer number of neurons available for a given task affect the computational capability of even a symbolic neural network.

The models discussed in this workshop highlight the necessity of incorporating biological data into theoretical models and computer simulations. Take, for example, the historical caricature of the grandmother cell, cited as evidence against local representations. When modeled as part of the entire visual system, the "grandmother cell" is usually allocated to the top slot in a strict hierarchy of levls. At each successive level, cells perform an increasing abstraction of the input stimulus and receive ever more convergent input from lower levels. The levels are assumed to interact in a strictly feedforward fashion, beginning with a retinal stimulus and culminating in the activity of a cell representing, say, grandmother viewed straight-on at eye level. This model is at best incomplete as an explanation of the visual system, even as an analogy, because it fails to incorporate critical anatomical and physiological data. These are data that illustrate the abundance of lateral and feedback connections (e.g., Felleman and Van Essen 1991; Rockland and Lund 1982). The possible functions of the numerous reciprocal (reentrant) pathways have been explored in several computational models (e.g., Finkel and Edelman 1989). The mere existence of feedback connections between visual areas suggests an anatomical conveyance for the top-down processes hypothesized to play a role in segmentation and object recognition.

Damasio's model of the visual system (see Damasio and Damasio, chapter 17) ties down the idea of a dynamic, distributed representation to specific cortical areas and connections. While based on anatomical and physiological data, it impacts on the study of the brain at a higher level, by predicting the behavioral consequences of lesions in specific brain areas. The phenomenology of object recognition described by Damasio's model includes evidence from prosopagnosiacs that correlates preserved levels of categorization within recognition with undamaged anatomical structures. A patient with bilateral mesial occipital lesions (including areas 18 and 19), sparing left-sided primary visual cortex, retains recognition of faces only at the basic-level. He can distinguish a face from a foot, but only rarely a man's face from a woman's, and he cannot recognize facial expressions at all (Tranel et al. 1988). Patients with lesions sparing extrastriate visual cortex on a least one side can recognize facial expressions and gender but cannot identify individual faces or certain man-made objects. A patient with a lesion sparing much of left-sided inferotemporal cortex and mesial occipital cortex on both sides can recognize some faces on a unique level, albeit slowly and by relying on strategies such as recognition-by-parts. Patients with bilateral lesions of inferotemporal cortex cannot recognize faces on a nonunique level but can identify facial expressions and gender. Thus, it seems that anterior temporal cortices are essential for recognizing the identity of faces, while early extrastriate cortices are required for distinguishing physical categories of faces.

Damasio's framework for object recognition (Damasio 1989) collects this and other evidence and combines it with neuroanatomical data on cortico-cortical connections. His model postulates that objects are represented by the coactivation of neurons distributed across many areas and levels of the visual, association, and limbic cortices, and linked by a network of feedforward and feedback connections that resist being labelled with a single direction of sequential processing. No single neuron or group of neurons in a single area has conferred upon it the entire structure of an entire object. In particular, neurons in so-called "integrative" areas such as inferotemporal cortex do not re-represent the elementary features registered by primary sensory areas. Thus the loss of inferotemporal cortex does not result in a breakdown of perception in the way that lesions to areas 18 and 19 may disrupt, say, stereopsis.

This framework is consistent with neurophysiological data on single neuron responses, since no cells have been discovered in monkey inferotemporal cortex, or elsewhere, that are specific for individual faces. The behavioral pattern of deficits supports the hypothesis derived from GRBF-like models of object recognition that individual faces may be encoded by the population activity of cells broadly tuned to particular views (Perrett et al. 1985) or to facial prototypes (Young and Yamane 1991).

These results underscore the theoretical and practical difficulties in determining the response specificities of neurons in the visual pathway. That prosopagnosiacs have similar deficits in the recognition of other objects suggests that extrapolating from faces to objects is valid and that cells specific for other objects may be found in inferotemporal cortex. "Hand" cells indeed were reported there even before "face" cells (Gross et al. 1972). But all the evidence suggests that cells with specificities as clear-cut as simple cells in primary visual cortex are not to be found in more anterior visual areas. Cells in these areas are more complex, at least in that they are far more difficult to excite. Analyses will have to be made very carefully of both stimulus properties and population responses. Indeed, given the anatomical connections between the inferotemporal cortex and limbic areas, it may be misguided to investigate the responses of an inferotemporal cell with purely visual stimuli (see e.g., Sakai and Miyashita 1991).

UNANSWERED QUESTIONS

Many of the models discussed in this workshop did indeed raise questions that could motivate new experiments. Some of the critical points put forward were:

1. Temporal synchrony for binding: The question arises of temporal resolution. Can the functional capabilities of individual cells support their group behavior? Because of saccades, microsaccades and drift, even a static image may send a rapidly changing signal within which different patterns must be distinguished by different temporal codes. Given the biophysical properties of neurons, is this feasible? To date only moving stimuli have been used in physiological

experiments to elicit temporal synchrony. These have an intrinsic temporal structure that could elicit temporal patterning of neuronal responses. It remains to be shown that temporal synchrony occurs in the binding of stationary features. Specific stimuli used to probe figure/ground discrimination abilities in humans have been proposed: for example, the noisy circle depicted by Ullman (this volume). Do the amplitudes of or temporal relations between responses of individual cells or cell groups alter when randomly scattered segments are rearranged to form a circular contour? How would the temporal synchrony model of binding account for the observed ability to group together features of a moving object, that would necessarily activate a continuously shifting set of cells? Can experimental evidence be found for temporal synchrony in the representation of such a stimulus? Specifically, is there a difference in correlation of activity between two cells stimulated simultaneously by two lines oriented in the same direction and moving in the same direction depending on whether the line segments have been moving coherently or not across previously stimulated receptive fields? On the psychophysical level, can the Gestalt-criteria-governed patterns of figure/ground discrimination be disrupted by an externally imposed temporal code?

2. 2D View Interpolation Schemes : These would predict that in an awake monkey learning to discriminate novel views of a wire-frame object, cells more or less broadly tuned to the 2D training views would be found, probably in inferotemporal cortex. We still have little idea of what actually happens during the 200 msec between stimulation of the retina and the initiation of a behavioral response, sufficient time for activity to travel all the way from retina to inferotemporal cortex and back to the earliest stages of cortical processing. For example, what does the stretch of cortex between V4 and inferotemporal cortex do? Following what guidelines might the neurophysiologist select the optimal stimuli for relatively unexplored visual areas? Given the impossibility of employing an infinite set of stimuli from simple bars to complex shapes, the neurophysiologist can never claim that any one stimulus is optimal for a given cell. In this new age of cell groups, what new information can the single cell recording reveal? Will our understanding of the visual system stagnate until new high-resolution technology for multiple-site recording appears?

One notion pervades almost all models of the visual system: that of a hierarchical structure. In computational vision, processes are relegated to "early" or "late" vision; in psychophysics, they are described as "low" or "high level." Physiology and anatomy have taught us that these distinctions make some sense: the retina is the first to receive the light signal, and cortical areas removed from it by a greater number of synapses tend to have more abstruse visual functions. But as "higher" visual areas are explored in more detail, the hierarchy that labels them may itself tumble to make way for a different sort of vision.

REFERENCES

Artola, A., S. Brocher, and W. Singer. 1990. Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature* **347**:69–72.

Barlow, H.B. 1972. Single units and sensation: A neuron doctrine for perceptual psychology? *Perception* **1**:371–394.

Bauer, K., and U. Krey. 1990. On learning and recognition of temporal pattern sequences. *Zeitschrift fur Physik B* **79**:461.

Biederman, I. 1987. Recognition by components: A theory of human image understanding. *Psychol. Rev.* **94**:115–147.

Bush, P. C., and R. J. Douglas. 1991. Synchronization of bursting action potential discharge in a model network of neocortical neurons. *Neural Comp.* **3**:19–30.

Crick, F., and C. Koch. 1990. Towards a neurobiological theory of consciousness. *Sem. Neurosci.* **2**:263–275.

Cumming, B., A. Hurlbert, E. Johnston, and A. Parker. 1991. Effects of texture and shading on the KDE. *Inv. Ophthalm. Vis. Sci. Suppl.* **32/4**:1277.

Damasio, A. R. 1989. The brain binds entities and events by multiregional activation from convergence zones. *Neural Comp.* **1**:123–132.

Damasio, A., H. Damasio, D. Tranel, and J. Brandt. 1990a. Neural regionalization of knowledge access: Preliminary evidence. *Symp. Quant. Biol.* **55**:1039–1047.

Damasio, A., D. Tranel, and H. Damasio. 1990b. Face agnosia and the neural substrates of memory. *Ann. Rev. Neurosci.* **13**:89–109.

Edelman, S., and D. Weinshall. 1991. A self-organizing multiple-view representation of 3D objects. *Biolog. Cybern.* **64**:209–219.

Egelhaaf, M. 1990. Spatial interactions in the fly visual system leading to selectivity for small-field motion. *Naturwiss.* **77**:182–185.

Engel, A.K., P. Konig, and W. Singer. 1991. Direct physiological evidence for scene segmentation by temporal coding. *Proc. Natl. Acad. Sci.* **88:**9136–9140.

Felleman, D.J., and D.C. Van Essen. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cortex* **1**:1–47.

Finkel, L.H., and G.M. Edelman. 1989. The integration of distributed cortical systems by reentry: A computer simulation of interactive functionally segregated visual areas. *J. Neurosci.* **9**:3188–3208.

Fischler, M., and R. Elshlager. 1973. The representation and matching of pictorial structures. *IEEE Trans. on Computers*, vol. 22.

Gerhardstein, P.C., and I. Biederman. 1991. 3D orientation invariance in visual object recognition. *Inv. Ophthalm. Vis. Sci. Suppl.* **32/4**:1181.

Ghose, G. M., and R. D. Freeman. 1990. Origins of oscillatory activity in the cat's visual cortex. *Soc. Neurosci. Abst.* **523**:4.

Gilbert, C., and T. Wiesel. 1989. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* **9**:2432–2442.

Gochin, P.M., E.K.Miller, C.G.Gross, and G.L. Gerstein. 1991. Functional interactions among neurons in inferior temporal cortex of the awake macaque. *Exp. Brain Res.* **84**:505–516.

Gray, C.M., P. König, A.K. Engel, and W. Singer. 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**:334–337.

Gray, C.M., and W. Singer. 1989. Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci.* **86**:1698–1702.

Gregory, R.L. 1970. The Intelligent Eye. London: Weidenfeld-Nicolson.

Gregory, R.L. 1986. Odd Perceptions. London: Routledge.

Grenander, U., Y. Chow, and D. Keenan. 1990. HANDS: A pattern theoretic study of biological shapes. New York: Springer.

Gross, C.G., C.E. Rocha-Miranda, and D.B. Bender. 1972. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* **35**:96–111..

Koch, C., and H. Schuster. 1992. A simple network showing burst synchronization without frequency-locking. *Neural Comp.* **4**, in press.

Koenderinck, J. 1984. The internal representation of solid shape and visual exploration. In: Sensory Experience, Adaptation and Perception, ed. Spillmann and Wooten. Lawrence, NJ: Erlbaum.

Koerner, E., and H. Boehme. 1991. Organization of an episodic knowledge base in a neural network architecture with parallel-sequential processing modes. In: Artificial Neural Networks, ed. Kohonen et.al., pp. 873–877. Amsterdam: Elsevier.

Kreiter, A.K., and W. Singer. 1991. Oscillatory neuronal activity in the superior temporal sulcus of macaque monkeys. *Soc. Neurosci. Abstr.* **17**:208.1.

Kumar, T., B.R. Beutter, and D.A. Glaser, 1992. Some effects of color on the perception of motion. *Perception* (submitted).

Murthy, V.N., and E.E. Fetz. 1991. Synchronized 25–35 Hz oscillations in sensorimotor cortex of awake monkeys. *Soc. Neurosci. Abstr.* **17**:126.11.

Nakayama, K., and S. Shimojo. 1990. Toward a neural understanding of visual surface representation. *Symp. Quant. Biol.* **55**:911–924.

Perrett, D. I., P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves. 1985. Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc. Lond. B* **223**:293–217.

Poggio, T., and S. Edelman. 1990. A network that learns to recognize 3D objects. *Nature* **343:**263.

Poggio, T., and F. Girosi. 1990. Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247:**978–982.

Poggio, T., and T. Vetter. 1992. Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries. Cambridge, MA: Artificial Intelligence Lab., MIT, AI Memo No. 1347.

Ramachandran, V.S. 1985. Guest editorial: The neurobiology of perception. *Perception* **14:**97–103.

Reichardt, W., T. Poggio, and K. Hansen. 1983. Figure-ground discrimination by relative movement in the visual system of the fly. Part II. *Biol. Cybern.* **46 (Suppl.)**:1–30.

Reiser, K. 1991. Learning persistent structure. Doctoral Thesis, Research Report 584, Hughes Aircraft Co., 3011 Malibu Canyon Road, Malibu, CA 90265.

Rockland, K.S., and J.S. Lund. 1982. Widespread periodic intrinsic connections in the tree shrew visual cortex. *Science* **215**:1532–1534.

Rosenblatt, A. 1961. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, DC: Spartan Books.

Sakai, K., and Y. Miyashita. 1991. Neural organization for the long-term memory of paired associates. *Nature* **354**:152–155.

Shepherd, R.N., and J. Metzler. 1971. Mental rotation of three-dimensional objects. *Science* **171**:701–703.

Sporns, O., J.A. Gally, G.N. Reeke, and G.M. Edelman. 1989. Reentrant signaling among simulated neuronal groups leads to coherency in their oscillatory activity. *Proc. Natl. Acad. Sci.* **86**:7265–7269.

Sporns, O., G. Tononi, and G.M. Edelman. 1991. Modeling perceptual grouping and figure-ground segregation by means of active reentrant connections. *Proc. Natl. Acad. Sci.* **88**:129–133.

Sporns, O., G. Tononi, and G.M. Edelman. 1992. Temporal correlations and integrative functions of the brain. *J. Cog. Neurosci.*, in press.

Swain, M.J., and D.H. Ballard. 1990. Indexing via color histograms. In: Proc. Internatl. Conf. on Computer Vision, pp. 390–393. Osaka: IEEE.

Tarr, M., and S. Pinker. 1989. Mental rotation and orientation-dependence in shape recognition. *Cog. Psych.* **21**:233–282.

Toyama, K, M. Kimura, and K. Tanaka. 1981a. Cross-correlation analysis of interneuronal connectivity in cat visual cortex. *J. Neurophysiol.* **46**:191–201.

Toyama, K, M. Kimura, and K. Tanaka. 1981b. Organization of cat visual cortex as investigated by cross-correlation technique. *J. Neurophysiol.* **46**:202–214.

Tranel, D., A.R. Damasio, and H. Damasio. 1988. Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology* **38**:690–696.

Treisman, A., and H. Schmidt. 1982. Illusory conjunctions in the perception of objects. *Cog. Psychol.* **14**:107–141.

Tso., D.Y., Gilbert, C.D., and T.N. Wiesel. 1986. Relationship between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *J. Neurosci.* **6**:1160–1170.

Turk, M., and A. Pentland. 1991. Eigenfaces for recognition. *J. Cog. Neurosci.* **3**:71–86.

Ullman, S., and R. Basri. 1990. Recognition by linear combination of models. Cambridge, MA: Artificial Intelligence Lab., MIT, AI Memo No. 1152.

Vaadia, E., E. Ahissar, H. Bergman, and Y. Lavner. 1991. Correlated activity of neurons: A neural code for higher brain functions? In: Neuronal Cooperativity, ed. by J. Krüger, pp. 249–279.

von der Malsburg, C. 1992. The correlation theory of brain function. *Cog. Neurosci.*, in press.

von der Malsburg, C., and W. Schneider. 1986. A neural cocktail-party processor. *Biolog. Cybern.* **54**:29–40.

Wertheimer, M. 1923. Untersuchungen zur Lehre von der Gestalt II. *Psych. Forsch.* **4**:301–350.

Young, M.P., and S. Yamane. 1992. An analysis at the population level of the processing of faces in the inferotemporal cortex. In : Brain Mechanisms of Perception and Memory: from Neuron to Behaviour, ed. L. Squire, M. Fukuda, and D. Perrett, New York: Oxford Univ. Press, in press.

Young, M. P. , K. Tanaka, and S. Yamane. 1992. On oscillating neuronal responses in the visual cortex of the monkey. *J. Neurophys.*, in press.

Yuille, A., D. Cohen, and P. Hallinan. 1989. Feature extraction from faces using deformable templates. *Proc. Comp. Vis. Pattern Recog.*, pp. 104–108.