1987 Short Course on Computational Neuroscience, Society for Neuroscience

# NEURAL NETWORK MODELS OF VISUAL PROCESSING

Terrence J. Sejnowski and Sidney R. Lehky

Department of Biophysics
Johns Hopkins University
Baltimore, MD 21218

## CONTENTS

# I. Introduction

## A. Levels of Analysis

There are many different scales at which the nervous system can be studied, from the molecular to the behavioral levels, including the synaptic, neuronal, local circuit, network, and systems levels. The understanding of neurons and their synaptic interactions is crucially important for studying networks of neurons. Neural networks may, however, have properties that can not be predicted from the properties of single neurons studied by themselves. The purpose of this chapter is to give a brief introduction to techniques in neural network modeling that may help in studying these emergent properties. Two examples will be presented from our own work to illustrate the general issues.

At present it is possible to model groups of up to about 1000 interacting neurons. This is sufficient to study small circuits such as those found in invertebrate ganglia and cortical columns. Emergent properties that require networks of neurons are expected to arise in such circuits, but current experimental techniques provide only poor information about how this is done. However, over the next decade new techniques, particularly recordings from multiple electrodes and optical recording with voltage and ion-sensitive dyes, are likely to provide insight into this level of information processing in neural networks. Neural network models could help in interpreting this data and in designing new experiments.

Two broad classes of neural network can be distinguished according to the role played in the network by single neurons. In some circuits, such as pattern generators in invertebrate ganglia, each neuron has a special role that may make it unique (Selverston, 1985). In contrast, a neuron in the retina of a vertebrate participates in circuits with many other similar neurons. The pattern of activity in the population of neurons is important in the function of the circuit. This chapter is concerned with the modeling of neural populations, and the primary focus is on neural circuits in the early stages of visual processing in mammals. Many of the techniques and general principles should also apply throughout the nervous system of vertebrates.

## B. Bottom-Up and Top-Down Strategies

As experimental data accumulate at the level of single neurons, more and more detailed models become possible that mimic progressively more closely the detailed processing of particular circuits. This approach, which might be called "bottom-up", is most useful when 1) The function of the circuit is already known and 2) the knowledge about the circuit is almost complete down to the biophysical level. In most parts of the vertebrate central nervous system we have only a vague notion about the function of circuits and information about the biophysical level is at best incomplete. One of the lessons learned from modeling invertebrate circuits is that a wiring diagram is not nearly enough to specify a circuit -- specific membrane properties are also crucial. Unfortunately, even the patterns of connectivity are uncertain in cerebral cortex.

Another approach to the network level is to start with a function such as a perceptual ability and design simplified neural circuits that can perform the function within the constraints of the state of knowledge. This could be called a "top-down" approach. One example is the Marr-Poggio (1976) model of binocular depth

perception, which demonstrated that a network of simplified model neurons could fuse random-dot stereograms (Julesz, 1960). These models can be simulated on a digital computer and their performance compared with psychophysical measurements. Unfortunately, these models are often too general to directly compare with physiological experiments, and at best they serve as a demonstration of one way that a problem can be solved.

Neither the top-down nor the bottom up approach is ideal -- what is needed is some approach that combines the strengths of both strategies. The model should be constrained by data at the neuronal level and should be informed by by a task-level analysis of the computational function of the system. Thus, constraints both from below and above should be incorporated into the model

The two examples from our own work given here illustrate the usefulness of a combined "outside-in" style of network modeling. They are at the level beyond that of modeling the details of identified neurons, but not so general that essential features such as response properties of single neurons can no longer be identified.

## II. Binocular Rivalry

This section will provide an example of a neural model applied to a particular aspect of binocular vision. Knowledge about both neurobiology and psychophysics will be incorporated into the model, each field providing a complementary set of information. From neurobiology we have information about the physical substrate of the system, the anatomy and the physiology of the neurons. Yet in considering the data collected at the level of single neurons, it is often difficult to establish any clear sense of what the system as a whole is doing. In this case it may be useful to step back and gain an overview of the operation of the intact, functioning system as determined by psychophysics.

The phenomenon that will be considered here is binocular rivalry. Rivalry occurs when unmatchable images are presented to the two eyes, such as vertical stripes to one eye and horizontal stripes to the other. In such as situation the visual system is thrown into oscillation, so that first the image from one eye is visible, and then the other, typically for a period of about one second. In general, the entire visual field does not oscillate in unison unless the rival stimuli are sufficiently small (less than $1^o$ in diameter), but rather it breaks up into a constantly changing mosaic of the incompatible images.

Binocular rivalry was chosen for modeling rather than stereopsis, which is the aspect of binocular vision that has received the most attention, both experimental and theoretical. Rivalry appears to be a simpler problem, while at the same time retaining sufficient complexity that studying it may yield interesting insights into the general organization of binocular vision. Rivalry also appears to be a problem of intermediate complexity in the sense that the issues can be formulated in terms relevant to the biological concerns of the experimental neurophysiologist as well as the global concerns of the psychophysicist.

### A. Experimental Data

A large body of psychophysical data related to rivalry has accumulated over the past century (O'Shea, 1983). Although the durations of alternating left and right dominance show statistical variation in length, the mean durations depend on the stimuli. The nature of this dependence is unusual. When the stimulus strength (contrast, for example) is increased to one eye, the duration of time for which the opposite eye is dominant decreases. It is therefore possible to independently vary the duration that each eye is dominant. Also, the oscillations follow a rectangular waveform. This is indicated by the observation that a spot of light flashed to the suppressed eye has its detectability reduced by a constant amount over the entire duration of the suppressed phase, interpreted as showing that the strength of suppression remains constant until being abruptly cut off.

Neurophysiological data on the matter is more sparse. Allman (unpublished data) has observed in the superficial layers of owl monkey primary visual cortex neural activity that switches on and off in synchrony with behavioral indications from the alert animal that an eye had undergone a transition from suppressed to dominant states. This is the only neurophysiological report of oscillations associated with rivalry. Varela and Singer (1987) have recorded from LGN relay cells of anesthetized cats exposed to rivalrous stimuli. They found that strong inhibition occured when stimuli to the two eyes were unmatched. This inhibition had a latency of hundreds of milliseconds, and was abolished by disruption of corticofugal inputs through ablation of the cortex. Although no oscillations were observed (possibly because of the anesthesia), these data suggest a binocular inhibitory process at an early stage of the visual system whose activity is related to the degree of correlation between images to the two eyes. Together with the results of Allman, these studies indicate that although rivalry has been chiefly a concern of psychophysics, it is feasible to approach the phenomenon with neurophysiological techniques. This may be one of the simplest experimental paradigms that could link conscious awareness of sensory stimuli with neural activity.

### B. Neural Network Model

By its very nature, binocular rivalry suggests some sort of reciprocal inhibitory linkage between signals from the left and right eyes prior to the site of binocular convergence. Reciprocal inhibition is common at all levels of the nervous system, from peripheral processing in the retina to visual cortex. The simplest neural network implementation of such as system is illustrated in Fig. 1, which shows the responses of the network under different stimulus conditions, discussed below (Lehky, 1987). The open circles represent excitatory neurons and the filled circles inhibitory neurons. There are two excitatory neurons, one from the left side and one from the right, which converge upon a binocular neuron. These two neurons receive sensory information along inputs from the periphery, indicated by the lines originating from the top of the diagram. Finally, there are two inhibitory neurons, each inhibiting the other side in feedback fashion. The model is incomplete, as will be discussed later, and is only valid for a small patch of the visual field.

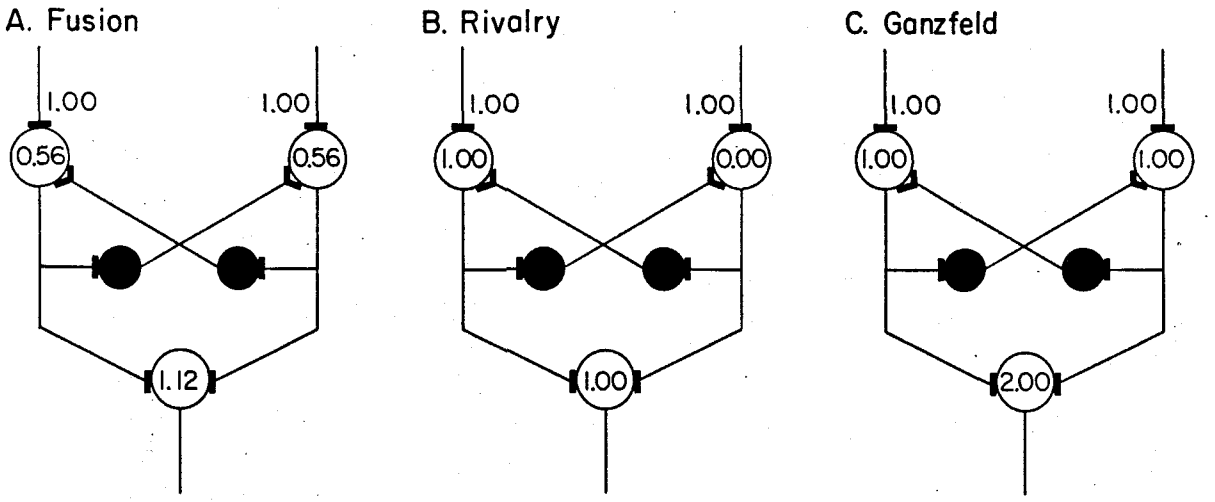



*Figure 1: Network model of binocular rivalry shown with three different stimulus conditions*

Rather than model actual neurons, the point of this network is to provide as simple a model as possible of the way in which oscillations could be produced by neural interactions. Oscillations of course imply a system whose state is changing as a function of time, or in other words, a dynamical system. The behavior of this system can be found using the mathematical methods of stability theory, which will give the conditions under which the system will assume various qualitative states (i.e. whether it is in an oscillating or non-oscillating state) without providing quantitative information about the state. When a stability theory analysis is performed on a reciprocal feedback inhibition model, it can be shown that two conditions are required for it to go into oscillation. The first is that the inhibition due to one side must be sufficiently strong when first established to switch off the other side. The second requirement is that the inhibition strength must then adapt downward so as to pass below some level to permit the other side to activate.

### C. Analog Electrical Circuit

In order to go beyond the simple qualitative analysis described above, it is necessary to simulate the behavior of the system. The dynamics of the system are governed by coupled nonlinear differential equations which are not analytically soluble. Therefore one must either solve them numerically on a computer, or find an equivalent physical system that is subject to the same equations but can be more easily studied.

For the rivalry model, the system that was chosen as an analog to the neural network (analog in the sense that its behavior shared the dynamical features of interest

with the neural network) was the electronic circuit shown in Figure 2. The circuit is called an astable multivibrator, and is essentially an oscillating flip-flop, whose behavior is described in many electronic textbooks, and which is easily built for oneself.

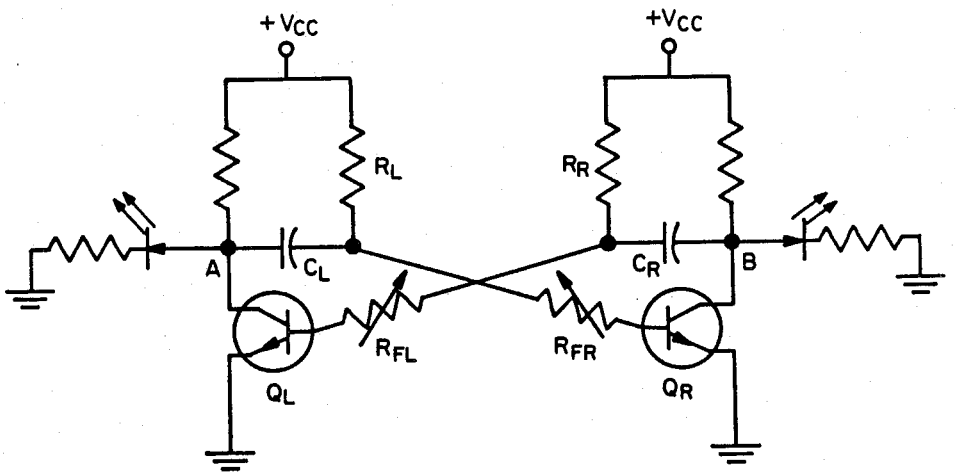



*Figure 2: Electrical analog model of binocular rivalry*

The analogy between this circuit and the neural network can be seen as we step through the operation of the circuit. The two transistors $Q_l$ and $Q_r$ represent the left and right excitatory neurons. The points labeled A and B in the circuit are equivalent to the outputs of those two neurons, and LED's were placed at those points to allow visual monitoring of the output voltages at those points. Nothing in the electronic circuit corresponds to the the binocular neuron. It would have been trivial to model a binocular neuron by including circuitry that linearly summed the voltages from points A and B, but that would not have added to our conceptual understanding of what was going on, and would have cluttered and complicated the circuit.

The left and right transistors are connected in a manner which may be described as reciprocal inhibition. The "inhibitory" pathway from $Q_l$ to $Q_r$ runs through $C_l$ and $R_{fr}$ to the base of $Q_r$, and analogously from point B to the base $Q_l$ on the other side. If the voltage at point A is high, that forces the voltage at point B to be close to zero. As the circuit oscillates, the voltages at points A and B alternately go high and low.

The slow charging of the capacitors $C_l$ and $C_r$ along the circuit path connecting the two transistors can be thought of as "adaptation" of inhibition, and this adaptation is necessary for the system to go into oscillation. When the voltage of capacitor $C_l$ on one side gradually charges up to the threshold of of transistor $Q_r$ on the other side, the system flips state. At this point the whole process starts over with the other

capacitor and transistor, and then so on, back and forth indefinitely as the system oscillates. (The transistors here effectively act as switches. That is to say, their transfer function between "input" base voltage and "output" collector voltage can be approximated by a very steep sigmoid.)

Finally, the values of the variable resistors $R_{fl}$ and $R_{fr}$ in the feedback pathway can be thought of as determining the strength of "inhibitory coupling" between the right and left sides. Small resistances correspond to strong inhibitory coupling and large resistances to weak coupling. As we said before, the strength of inhibitory coupling is one of the factors that will determine whether a system with reciprocal feedback inhibition will oscillate or not. This can easily be demonstrated in this circuit, for as one increases the values of the feedback resistance (decrease inhibitory coupling) by turning the knob of a potentiometer, a particular point is reached at which the oscillations suddenly stop. Instead of each LED being alternatingly fully lit or completely dark, both are now on simultaneously, but at some intermediate level of brightness. In the language of dynamical systems theory this is called a bifurcation point, a point at which a discontinuous change in the behavior of the system (from oscillating to non-oscillating) occurs as one of the system variables (feedback resistance) is continuously changed.

So now we can compare the behavior of this astable multivibrator circuit with that of binocular rivalry. One similarity is that the astable multivibrator produces rectangular oscillations. In addition, the duration of time that the left or right side is "dominant" can be varied independently by changing the value of a parameter (time constant for charging the capacitor) on the opposite side. This is analogous to the behavior of rivalry, if ones substitutes "change contrast" for "change capacitor (adaptation) time constant". Although it will not be discussed in detail here, the statistical distribution of the durations of dominance and suppression can also be replicated by the system under consideration here by adding random noise to the model (Lehky, 1987). Finally, the electronic circuit passes from an oscillating to non-oscillating states as the value of the feedback resistances (strength of inhibitory coupling) is varied. In the binocular visual system, we know that the behavior goes from oscillating (rivalry) to non-oscillating (fusion) as the correlation of images presented to the two eyes is varied.

This last point brings up an important physiological prediction of the astable multivibrator model, which is that not only is there reciprocal feedback inhibition prior to the site of binocular convergence, but that this inhibition involves synapses whose strength is affected by the degree of correlation between the left and right images. High correlation would lead to weak inhibitory coupling across the synapses, and low correlation would lead to strong inhibitory coupling. Furthermore, it has already been mentioned that changing stimulus strength (contrast) in rivalry and changing the adaptation time constant in the astable multivibrator circuit have analogous effects. Therefore, a second prediction would be that the binocular reciprocal inhibition postulated here shows adaptation whose time constant depends on contrast. In other words, adaptation occurs at a faster rate when contrast is increased.

It is important to point out here that the model as presented here is incomplete, since it just says that the degree of correlation between the left and right images affects the strength of inhibitory coupling, but does not give any mechanism by which

this may occur. Furthermore, it does not consider the spatial patterns occurring in binocular rivalry, which are presumably mediated by various lateral interactions.

Going back to Figure 1, the activities within the neural network as postulated by the model are shown for binocular fusion and rivalry. (The values for fusion are in fact taken from an earlier model (Lehky, 1983) which considered psychophysical results about how various luminances presented to the two eyes combine to form the perception of binocular brightness). The last diagram in the figure shows some psychophysical data from Bolanowski presented at the Neuroscience meeting in 1986, which shows that when no contours are present (Ganzfeld conditions) the binocular reciprocal inhibition disappears, so that luminances presented to the two eyes add up linearly to form the percept of binocular brightness. This is included here in support of a central point of the modeling, that the strength of inhibitory coupling in binocular vision is a function of the spatial patterns presented to the two eyes. The strength of inhibitory coupling for the three conditions can be ordered as follows: uncorrelated contour (rivalry) > correlated contour (fusion) > no contour (Ganzfeld).

A functional model such as the one presented here cannot define anatomical location. However, the physiological data of both Allman and of Varela and Singer suggest that the processes under consideration here are already occurring at an early stage (primary visual cortex). We find the proposal of Varela and Singer that inhibitory circuits in the LGN are gating signals to the cortex dependent on the correlation between the left and right images at the cortical level to be attractive. Certainly the precise binocular layering found in the LGN points to an important role in binocular vision, beyond the inadequate notion of the LGN as the recipient of a rather amorphous set of modulatory influences from the brainstem, or the simple notion of the LGN as a "relay center".

In conclusion, although the model presented here was at a much simplified level, it both fits the available data well and suggests new lines of experimental investigation that might not have been otherwise considered.

## III. Computing Curvature from Shaded Images

In the previous example intuition was coupled with knowledge of anatomy and physiology of the visual system to arrive at a plausible network model of binocular rivalry. The network is a small, but important part of a larger, more complex system. As processing is traced into visual cortex, it becomes more difficult to find plausible network models based on intuition. Recently, a new constructive technique has been discovered for designing layered networks that can perform specified transformations. In this section we describe the technique and its application to a problem in the visual processing of continuous-tone images.

### A. Interpreting Experimental Data

Ever since Hubel and Wiesel (1962) first reported that single neurons in the cat visual cortex respond better to oriented bars of light and to dark/light edges than to spots of light, it was generally assumed that the function of these neurons was to detect the boundaries of objects in the world. According to this view, a single cortical neuron could be thought of as detecting an edge with a particular orientation at a particular location in the visual field. However, it has also been suggested that cortical

cells should also be characterized according to their spatial frequency response function. It is by no means obvious how to infer the function of an oriented cortical cell from its response properties. This is a general problem that arises throughout the nervous system.

Boundaries of objects are rare in images, yet the majority of cells in visual cortex respond preferentially to oriented bars and edges. These cells also respond, though less vigorously, to texture and gradually changing spatial gradients of light. Do these partially-activated cells also carry useful information, and if so what function can they serve?

One of the primary properties of a surface is its curvature. Some surfaces are flat and have no intrinsic curvature, but others, like cylinders and balls, are curved. The curvature of a surface along a line through the surface can be computed by measuring the second derivative of the tangent vector to the surface along the line. The principal curvatures at a point on the surface are defined as the maximum and minimum curvatures, and these are always along lines that meet at right angles. The principal curvatures are parameters that provide information about the shape of a surface, and they have the additional advantage of being independent of the local coordinate system. Hence it would be helpful to have a way of estimating principal curvatures directly from the shading information in an image.

One of the problems with extracting the principal curvatures from an image is that the shading of a surface depends on many factors, such as the direction of illumination and the orientation of the surface relative to the viewer. Can information about the curvature of a surface be extracted locally from an image without additional information about the source of illumination (Pentland, 1984)?

### B. Constructing a Layered Network Model

Analytical techniques have been applied to the problem of computing the shape of a surface from its shading in computer vision (Ikeuchi & Horn, 1981), but this approach relies on biologically implausible assumptions and is difficult to reconcile with the nervous system. The approach that we have taken is to start with information that the visual cortex is known to receive and to transform this information through several layers of processing such that in the final output layer information is explicitly available about the curvature parameters of the surface that generated the image. Until recently it was not obvious how one would go about constructing a network to solve a problem such as the one proposed above. New techniques have been developed for constructing the network by training it with examples. The back-propagation algorithm (Rumelhart et al, 1986) that was used for our problem is described in the appendix.

Briefly, we generated a set of 2000 images of parabolic surfaces, each differing in the direction of the light source, the magnitudes of the principal curvatures, the direction of the maximum curvature, and location of the center of the surface. The reflectance function was assumed to be Lambertian (a matte reflectance). These images served as the training corpus for the neural network learning algorithm. The specific task was to develop a network which extracted the principal curvatures and their orientations independent of the illumination direction, and the precise location of the surface patch within the overall input receptive field of the network.

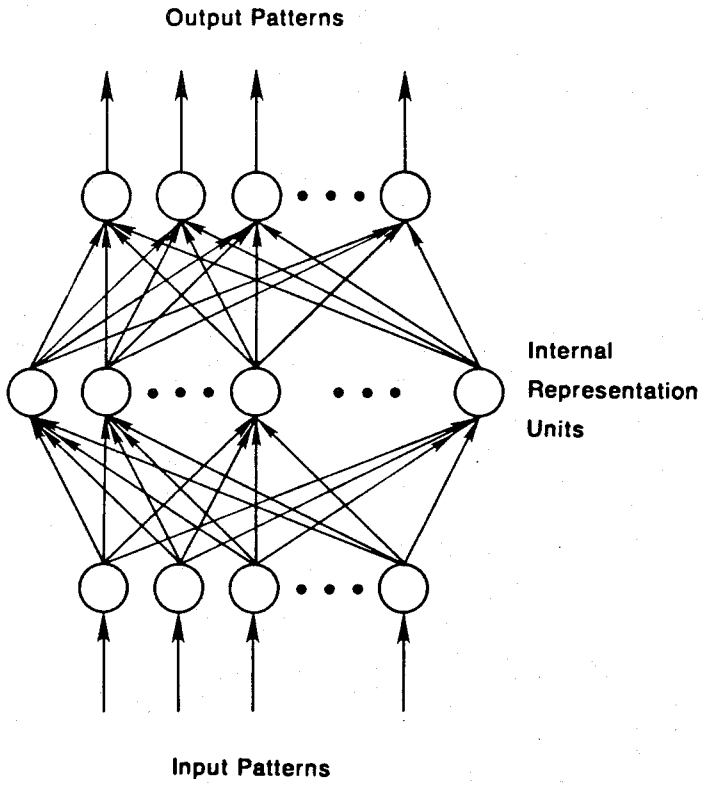
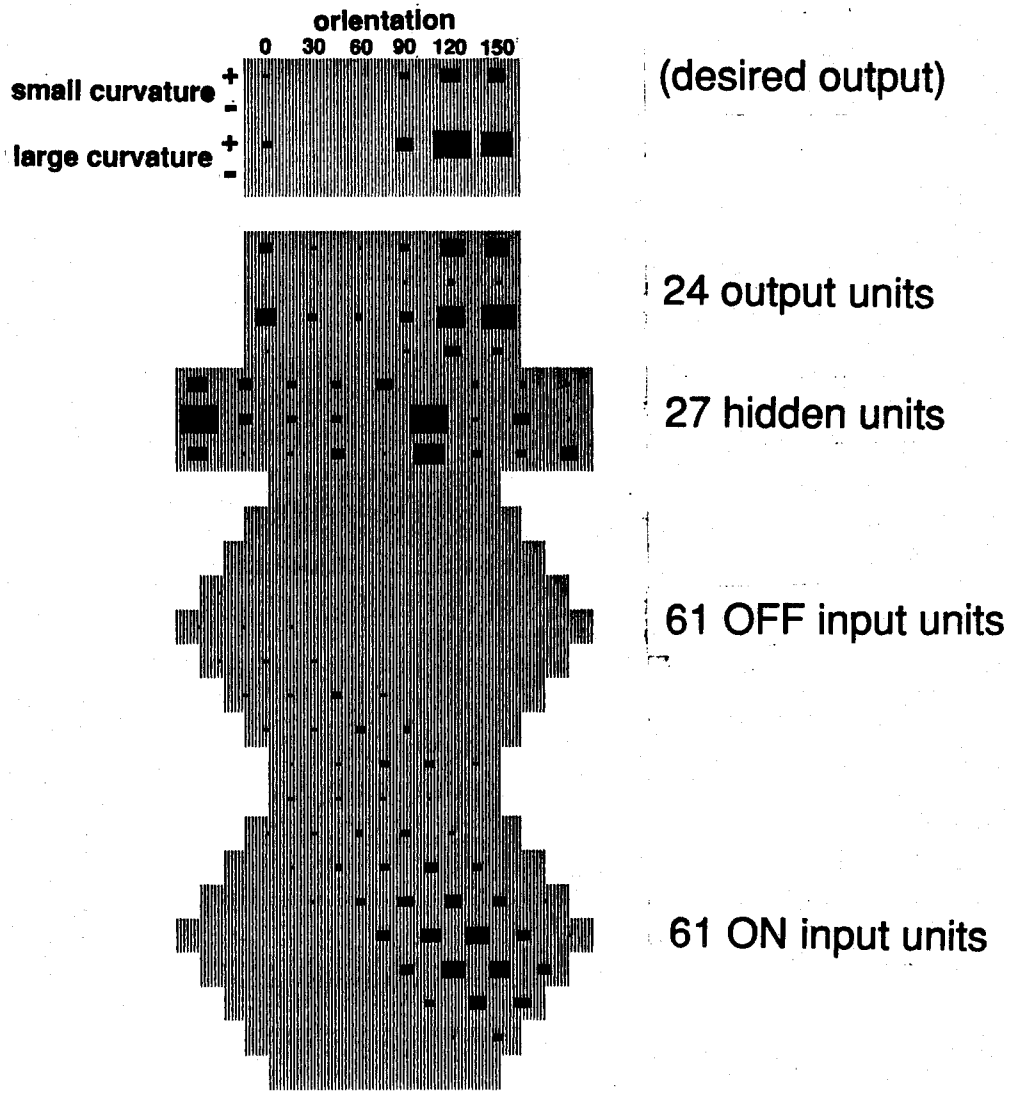



*Figure 3: Network model for shape from shading.*

## Response of network to an image



There are two sources of potential ambiguity in the data. This is because the signs of the two calculated curvatures (concave or convex) depend on whether the illumination is presumed to come from one side or the other. In this study we assumed that the light source is always from above, and that both curvatures have the same sign. There is in fact some evidence that the human visual system assumes that the light source is above the scene. It may be that the second assumption could be eliminated if the model were forced to come up with a self-consistent description of a larger, more complicated surface. The model as it now stands is exposed to only a small patch of surface in isolation, in essence like viewing the world through a long narrow tube. In terms of modeling visual cortex, the network only provides for local connectivity within a single column, and no provisions are made for interactions between columns. Eventually, an array of interacting networks could be constructed that resolve ambiguity from global information.

The particular network we used had three layers, an input unit layer, an output unit layer, and an intermediate hidden unit layer. This organization can be seen in Fig. 3, which shows the response of the fully developed network to a typical input image. The two hexagonal regions at the bottom represent responses of the 122 input units. The 27 units in the hidden layer are represented by the 3x9 rectangular array above the hexagons. Finally, the 24 units in the output layer is represented by the 4x6 array at the top. In all cases, area of the black squares is proportional to the activity of a particular unit. These three layers will be more fully described below.

The properties of the receptive fields for the input and output layers were given to the network, based on what operations we wanted the network to perform, as well as being constrained by biological plausibility. Through the learning algorithm, the network proceeded to develop receptive fields for the hidden units. These hidden unit receptive fields essentially act as a mapping, or transform, which converts the inputs to the desired outputs. Before further considering the hidden units and their receptive fields, we describe the receptive fields of the input and output units.

The input layer consisted of two hexagonal arrays of units, called the ON units and OFF units. These two arrays were superimposed on each other, so that each point of the image was sampled by both an ON unit and an OFF unit. Each of these arrays consisted of 61 units, for a total of 122 units in the input layer. (Sixty-one happens to come out to an even number on an hexagonal array.) The receptive field of each input unit was the Laplacian of a two dimensional gaussian, or in other words the classic circularly-symmetric center-surround receptive field found in the retina and LGN. This receptive field organization is illustrated in Fig. 4, which also shows that the receptive fields of the input units were extensively overlapped. The receptive fields of ON and OFF units were both of the same shape, but had opposite polarities. Responses of the input units to an image were determined simply by convolving the image with the units' receptive fields.

Besides being biologically plausible, choosing these particular input receptive fields was advantageous from a computational view for the particular problem at hand. Specifically, the responses of the of these center-surround receptive fields, acting as second derivative operators, tended to compensate for changes in appearance in the object arising from illumination coming in different directions.

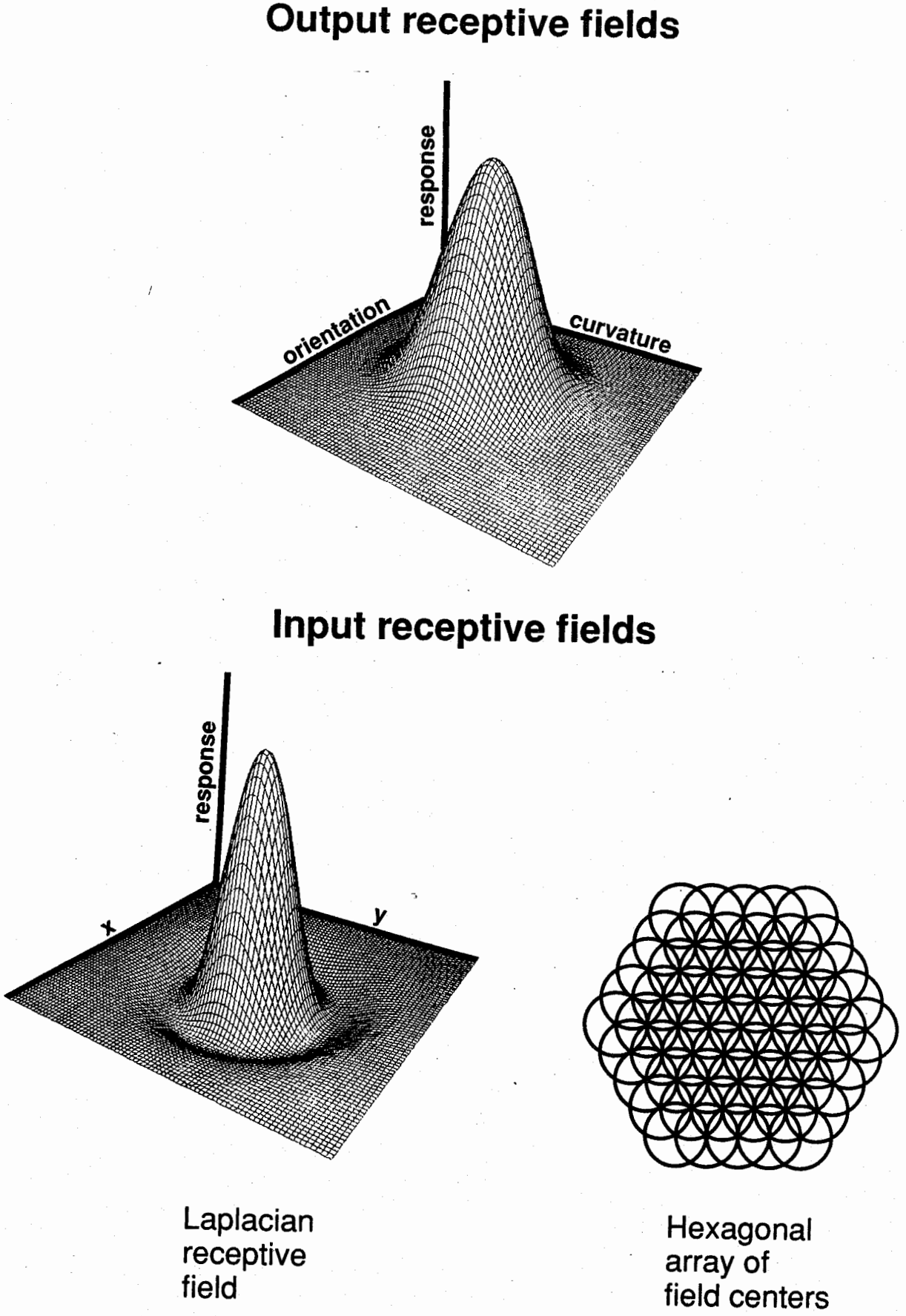



*Figure 4: Input and output receptive fields used in the network model of shape from shading.*

Moving to the output units, their receptive fields are also illustrated in Fig. 4. The figure shows that they give a graded response which is a function of both the value of the principal curvatures as well as their orientations. (Recall that these are the two parameters we are trying to extract from the images. It should also be noted that the curvature axis is on a logarithmic scale.) This sort of multidimensional response is typical of those found in cells of the cortex, although cells responding specifically to curvature have not yet been demonstrated.

However, the problem with having a nonmonotonic, multidimensional response is that the signal from the unit is ambiguous. There are an infinite number of combinations of curvature and orientation that give an identical response. The way to solve this ambiguity is to have the desired value represented in a distributed fashion, by the joint activity of a population of such broadly tuned units in which the units have overlapping receptive fields in the relevant parameter space (in this case curvature and orientation).

The most familiar example of this kind of distributed representation is found in color vision. The responses of any one of the three broadly tuned color receptors is ambiguous, but the relative activities of all three allow one precisely to discriminate a very large number of colors. Note the economy of this form of encoding; it is possible to form fine discriminations with only a very small number of coarsely-tuned units, as opposed to requiring a large number of narrowly-tuned, nonoverlaping units (Hinton, McClelland & Rumelhart, 1986). The output representation of parameters in the model under consideration here will follow the coarse tuning approach.

With that introduction to distributed representations, we can now examine the actual output representation used here, as illustrated in Fig. 3. Again, the output units are represented by the 4x6 array at the top of the figure. As one moves horizontally along a row, the output units are tuned to different peak orientations. For the image used in this example, the principal orientation was 130 degrees. One can see from the size of the black squares for the output units along a row that the largest responses come from units who have peak responses close to 130 degrees, and responses drop off as orientation tuning moves away from that value. The orientation value specified by any one unit is ambiguous, but the joint activities serve to precisely define orientation.

Moving vertically from top to bottom, the four rows represent different curvatures. The top two rows represent the value of the small principal curvature, and the bottom two rows represent the large principal curvature. (As was described above, there are two principal curvatures, and in general they will not be equal.) Within each of those two pairs of rows, the top row responds if the curvature is positive (convex surface) and the bottom one responds if the curvature is negative (concave surface). However, in the curvature domain, unlike the orientation domain, we have a set of non-overlapping tuning curves, and therefore curvature is not well represented in the model in its present state. (For both the large and small curvatures the peak curvature tunings are at +8 and -8, which is far enough apart that they don't overlap.)

The way to remedy the situation is to introduce a greater number of output units so that the curvature domain is sampled more densely, and the curvature tuning curves overlap. Alternatively, it is possible to sample the input image at different spatial scales (i.e. to have repeatedly convolved the input image with center-surround input

receptive fields with different diameters, instead of a single diameter as has been done). Both of these approaches are very time-consuming computationally, and neither has been implemented. This completes the description of the receptive fields of the input and output units.

### C. Analyzing the Hidden Units

The back-propagation learning algorithm described in the appendix was used to train a network with 27 hidden units. Briefly, the network was started with connection strengths between layers that had small random values (needed to allow each hidden unit to specialize on a different part of the problem). The 2,000 input patterns were presented to the network one at a time, and after each presentation the connection strengths were changed slightly to make the values of the units on the output layer compare more closely with the desired output values. Around 40,000 trials were required for the network to develop the connection strengths shown in Fig. 5, which was able to produce outputs that had an average correlation with the correct values of 0.9.

Each of the 27 hidden units is shown in Fig. 5 as a gray figure within which connection strengths are represented as black and white squares of varying size. The white square are excitatory weights, the black squares are inhibitory weights, and the area of the square is proportional to the magnitude of the weight. The two hexagonal arrays on the bottom of each figure shows the connections to the ON-center and OFF-center input arrays, and the rectangular arrays at the top are the connections to the units in the output layer. In addition, the value of the bias, or negative threshold, is shown in the upper left corner of the hidden unit.

The input arrays can be interpreted as the receptive fields of the units, insofar as spots or bars of light falling on the regions with excitatory inputs will tend to activate the unit. It can be seen that most of the hidden units had oriented receptive fields, similar to the pattern found in simple cells in visual cortex. However, two broad classes of hidden units could be distinguished by the pattern of connections with the output units. The connection strengths of the hidden units tended to discriminate for direction of the maximum principal curvature (vertical columns) or the magnitudes of the principal curvatures (horizontal rows), but not both. For example, the unit in the bottom right corner of Fig. 5 had strong inhibition for positive curvatures, but had little discrimination for orientation. Conversely, the unit in the bottom left corner had marked orientation preference, but little curvature discrimination. In addition, a few units, such as the one in the top left corner, had circularly-symmetric receptive fields and discriminated between the large and small curvatures.

The function of each hidden unit can be inferred from knowledge of the coding scheme used in the output layer and the pattern of the connection strengths between the hidden unit and the output units. The units that have strong output orientation preference are providing information primarily about the direction of maximum curvature. The units that discriminate positive from negative curvatures are responsible for providing information about whether the surface is convex or concave. The unit with a circularly-symmetric receptive field appears to contain information about the ratios for the principal curvatures. It should be noted, however, that the receptive fields of some hidden units were somewhat irregular, and that combinations of units might be

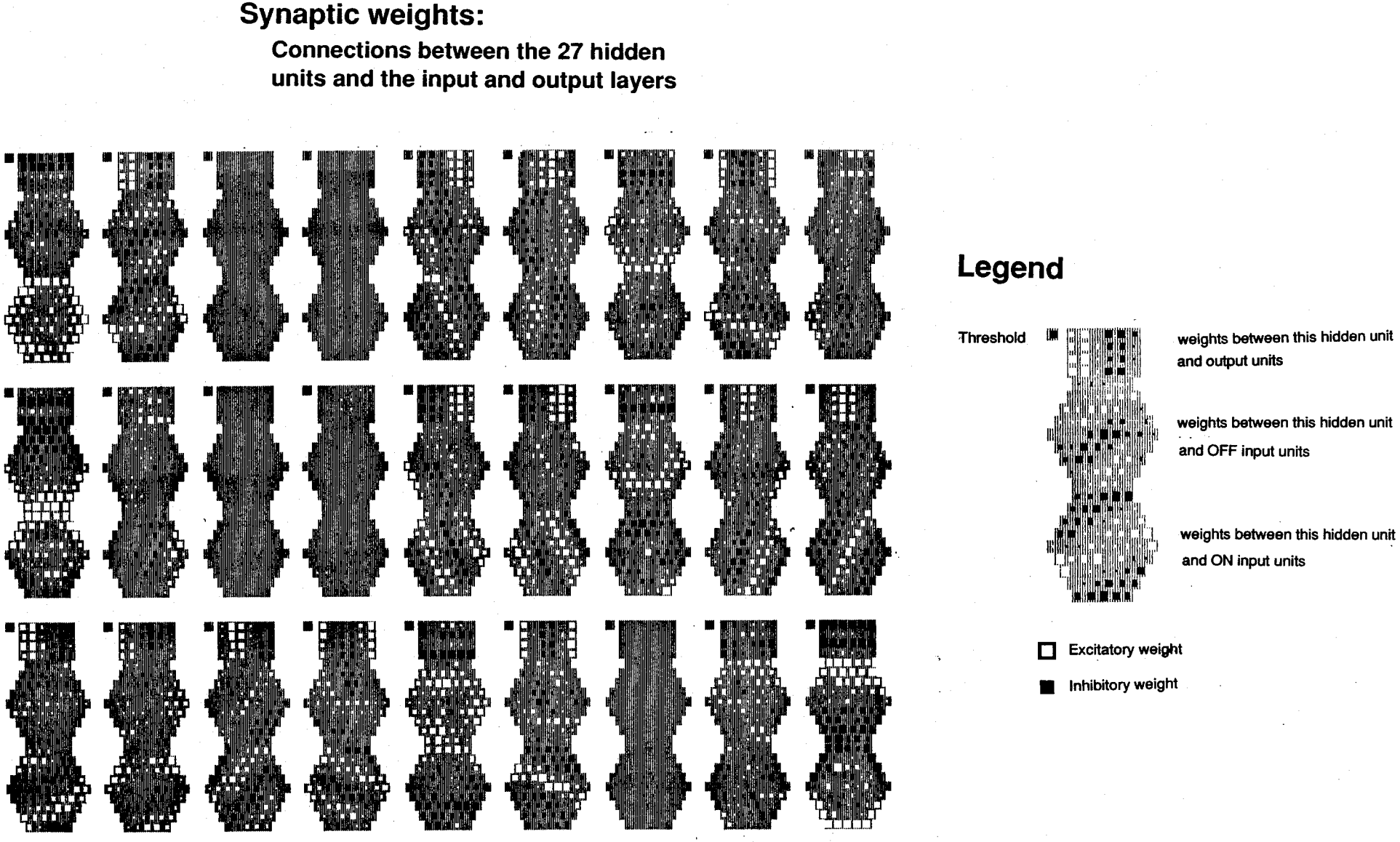



*Figure 5: Synaptic connection strengths for the hidden units in the shape-from-shading network.*

working together to provide information that is not apparent from examining single units.

The results of the learning procedure shown in Fig. 5 was representative of many training runs, each started with a different set of random weights. Similar patterns of receptive fields were always found, and the same three types of units could be found by examing the connections to the output units. However, there was variation in the details of the receptive fields, and in the number of units that did not develop any pattern of connections. It appears that only about 20 hidden units are needed to achieve the maximum performance, which was always the same average correlation of 0.9. The extra hidden units always undergo "cell death", since they serve no useful role in the network and can be eliminated without changing the performance.

The receptive field properties of the output cells were explored with bars of light that were varied in position, orientation, width and length. As expected, all of the output units had tuning curves for orientation, but unlike the hidden units that had identifiable excitatory and inhibitory subfields, the output cells were uniform in their response to the best oriented bar across the extent of the receptive field. This was a consequence of converging inputs from hidden units that had the same orientation preference but with different phases, as for the pair of units in the center of the middle row of units in Fig. 5. In addition to the orientation tuning, the output units were also sensitive to the length of the bar, and exhibited strong end-stopped inhibition of their responses. These are both properties of a subset of the cells in primary visual cortex with complex receptive fields.

The properties of the units on the hidden and output layers are consistent with the physiological properties of single units that have been found in primary visual cortex (Hubel & Wiesel, 1982). The network model provides an alternative interpretation of these properties, that they can be used to detect shape from shading rather than edges. The information contained in the shaded portions of objects in images can partially activate the simple and complex cells in visual cortex, and these responses can be used to extract curvature parameters from the image. It might prove interesting to test the response properties of cortical cells with curved surfaces similar to those used here.

## IV. Conclusions

### A. What Makes a Good Model?

It is clear that the present generation neural models cannot begin to reflect the complexity of the real nervous system. From anatomy and physiology we know that the visual system is a tangled web of multiple inputs, feedback loops, and lateral interaction, and that moreover, each unit within that web is a complex entity in itself, with the various nonlinear temporal and spatial integrative aspects of the dendritic tree being one part of that complexity. It is from a confrontation with this complexity, from a desire to see some pattern to it, that one is led to attempt the extraction of essential features and incorporate them into a simple model.

This leads to the most difficult part in constructing a model, which is to decide what is an "essential feature" and what is "simple". These are ultimately matters of intuition and judgement, although of course the choices are related to the types of

questions being asked (whether it concerns psychological phenomena or biophysical problems). At one extreme one could attempt to incorporate everything that is known about the nervous system into the model. The problem here is that the result is likely to be a model as incomprehensible as the nervous system itself. Although such a model may mimic the brain, it is debatable whether any true understanding has been achieved, since one remains unable to distinguish the relevant from the irrelevant. Going in the other extreme, the problems with oversimplification are obvious.

By what criterion can we decide whether a neural model is a good one or a bad one? One answer that is appealing at this stage in the development of neuroscience is that a good model is one that suggests new and promising lines of experimentation. In this spirit a neural model should be viewed on a provisional framework for organizing one's thinking about the nervous system. It doesn't matter if there are embarrassing gaps in the range of prediction the model offers, or that some assumptions seem a bit unrealistic. It doesn't matter if the details of the model are eventually proved wrong, for in the long run all models are wrong. As long as there is some kernel of truth that leads to new ways of thinking that prove productive, the model will have served its purpose. And while a model may be disproved by the very paths of inquiry that it engendered, in those same paths are the seeds of the next generation of models. A corollary to all this is that a very elaborate and sophisticated model that does not translate well into an experimental program is a sterile exercise when compared with something rougher that does translate readily.

### B. Binocular Rivalry

The proposed network model of rivalry is intended as a functional description rather than a detailed model, and the "equivalent circuit" in the model may eventually translate into a much more complex circuit in the real nervous system. For example, the individual neurons in the model network are likely to represent populations of real neurons that have some functional properties in common. Another limitation of the model is the difficulty in precisely specifying where in the system the neurons should be found. However, the model suggests places to look and predictions for what might be found there.

The model suggests that it is worthwhile to investigate sites of reciprocal inhibition prior to the site of binocular convergence. The suggestion depends on both the psychophysical data and knowledge of mechanisms that are known to exist in the nervous system in the areas of interest. All of this provides experimentalists some justification to embark on a line of investigation that they might not otherwise have thought interesting.

### C. Shape from Shading

In the proposed network for extracting curvature parameters from shaded images, the receptive field properties needed at the intermediate level of hidden units were similar to the properties of simple cells in primary visual cortex. These properties were not determined by the intuition of the model builder, but by the gradient descent procedure used by the learning algorithm. The modeler had only to specify the function that the network was required to perform by giving examples of inputs and the desired outputs. The network that was created was able to compute good curvature

parameters for new images as well as the ones it was trained on.

The network demonstrates that detecting bounding contours is not the only possible function of cells with simple receptive field properties in visual cortex. Sharp boundaries were excluded from the training set so that the only cues that network had available to compute the curvatures were in the shading. The information about the curvature parameters of a particular image were contained in the distributed pattern of activity in hidden unit layer.

The function of a single cell in the hidden layer was only revealed when its outputs were examined. This might be called the "projective field" of the unit, in analogy with the receptive field, which is defined for the pattern of inputs that drive it. The projective field provides the missing information needed to interpret the computational role of the unit in the network, and this can be inferred only indirectly by examining the next stage of processing. The sensitivity of simple and complex cells in visual cortex to the simple shaded patterns used in this study could be tested and compared with the network.

The prospect that some neurons in visual cortex are computing information about the curvature of objects is greatly strengthened by the psychophysical measurements demonstrating that humans can use shading information to estimate local surface normals in images (Mingolla & Todd, 1986). Humans, however, are only able to extract an approximate estimate of the curvature from local shading information, and it will be interesting to compare the accuracy of the network with that of humans under controlled psychophysical conditions. It is also clear that humans use other cues to estimate curvatures, such as the outline of bounding contours. This suggests that the network presented here should be considered only a small part of a much more complex system that uses multiple cues. It may be possible to study each of these cues separately and to propose network models for each of them that eventually can be combined.

### D. Other Applications

The two examples given in this chapter are specific examples of how the techniques of computer simulation and learning algorithms can be used to model neural circuits and networks. These techniques have much wider application to other problems. For example, Richard Anderson and David Zipser (unpublished) have recently applied back-propagation learning to help interpret the response properties of cells in parietal cortex. Recording from parietal cells indicate that some neurons in parietal cortex have receptive fields that are modulated by eye position. They used a three-layer network similar to the one described above to show that the same receptive field properties emerged when a network for trained to convert the retinal coordinates of an object to head-centered coordinates given information about the eye position.

Neural network modeling is still at an early stage of development, but it is already clear that new principles are emerging concerning the representation of information in neural populations, and transformations that are possible with these coding schemes. For example, Georgopoulos (1986) has shown that in motor cortex information about the intended direction of arm movement is distributed in populations of neurons that are broadly tuned to the direction. Network modeling can serve as a useful technique in conjunction with experiments designed to explore these principles in

sensory and motor systems.

## V. Appendix: Back-Propagation Learning Algorithm

The properties of the nonlinear processing units used in the model network in the curvature problem include i) the integration of diverse low-accuracy excitatory and inhibitory signals arriving from other units, ii) an output signal that is a nonlinear transformation of the total integrated input, including a threshold, and iii) a complex pattern of interconnectivity. The output of a neuron is a nonlinear function of the weighted sum of its inputs, and this can be approximated by the output function shown in Fig. 6.

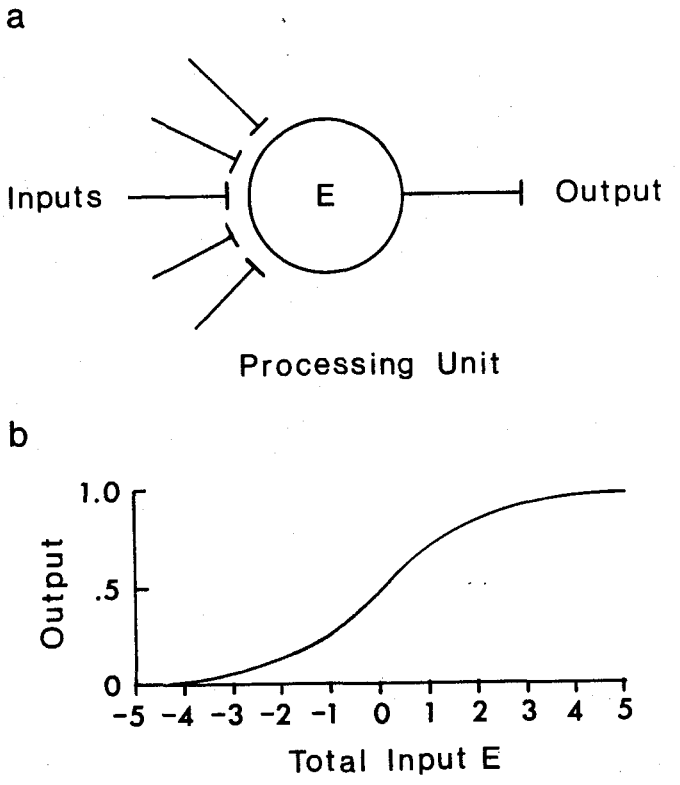



*Figure 6: Input-output function for the model neuron used in the back-propagation algorithm.*

This function has a sigmoid shape: it monotonically increases with input, it is 0 if the input is very negative, and it asymptotically approaches 1 as the input becomes large. This roughly describes the firing rate of a neuron as a function of its integrated input: if the input is below threshold there is no output, the firing rate increases with the input, and it saturates at a maximum firing rate. The behavior of the network does not depend critically on the details of the sigmoid function, but the one that we used is given by

$$s_i = P(E_i) = \frac{1}{1 + e^{-E_i}}, \tag{1}$$

where $s_i$ is the output of the $i$th unit and the total input $E_i$ is

$$E_i = \sum_j w_{ij} s_j, \tag{2}$$

where $w_{ij}$ is the weight from the $j$th to the $i$th unit. The weights can have positive or negative real values, representing an excitatory or inhibitory influence.

In addition to the weights connecting them, each unit also has a threshold. In some learning algorithms the thresholds can also vary. To make the notation uniform, the threshold was implemented as an ordinary weight from a special unit, called the true unit, that always had an output value of 1. This fixed bias acts like a threshold whose value is the negative of the weight.

Back-propagation is an error-correcting learning procedure that was introduced by Rumelhart, Hinton and Williams (1986). It works on networks with multilayered feed-forward architectures. There may be direct connections between the input layer and the output layer as well as through the hidden units. A superscript will be used to denote the layer for each unit, so that $s_i^{(n)}$ is the $i$th unit on the $n$th layer. The final, output layer is designated the $N$th layer.

The first step is to compute the output of the network for a given input using the procedure described above on successive layers. The goal of the learning procedure is to minimize the average squared error between the computed values of the output units and the correct pattern, $s_i^*$, provided by a teacher:

$$Error = \sum_{i=1}^{J} (s_i^* - s_i^{(N)})^2, \tag{3}$$

where J is the number of units in the output layer. This is accomplished by first computing the error gradient on the output layer:

$$\delta_i^{(N)} = (s_i^* - s_i^{(N)}) P'(E_i^{(N)}), \tag{4}$$

and then propagating it backwards through the network, layer by layer:

$$\delta_i^{(n)} = \sum_j \delta_j^{(n+1)} w_{ji}^{(n)} P'(E_i^{(n)}), \tag{5}$$

where $P'(E_i)$ is the first derivative of the function $P(E_i)$ in Fig. 6.

These gradients are the directions in which each weight should be altered to reduce the error for a particular item. To reduce the average error for all the input patterns, these gradients must be averaged over all the training patterns before updating the weights. In practice, it is sufficient to average over several inputs before updating the weights. Another method is to compute a running average of the gradient with an exponentially decaying filter:

$$\Delta w_{ij}^{(n)}(u+1) = \alpha \Delta w_{ij}^{(n)}(u) + (1 - \alpha) \delta_i^{(n+1)} s_j^{(n)}, \tag{6}$$

where $\alpha$ is a smoothing parameter (typically 0.9) and $u$ is the number of input patterns presented. The smoothed weight gradients $\Delta w_{ij}{}^{(n)}(u)$ can then be used to update the weights:

$$w_{ij}{}^{(n)}(t+1) = w_{ij}{}^{(n)}(t) + \varepsilon \Delta w_{ij}{}^{(n)} , \qquad (7)$$

where the $t$ is the number of weight updates and $\varepsilon$ is the learning rate (typically 1.0). The error signal was back-propagated only when the difference between the actual and desired values of the outputs was greater than a margin (typically 0.1). This insures that the network does not overlearn on inputs that it is already getting correct. This learning algorithm can be generalized to networks with feedback connections (Rumelhart et al., 1986), but this extension will not be discussed further.

The definitions of the learning parameters here are somewhat different from those in Rumelhart, Hinton and Williams (1986). In the original algorithm $\varepsilon$ is used rather than $(1-\alpha)$ in Eq. 6. Our parameter $\alpha$ is used to smooth the gradient in a way that is independent of the learning rate, $\varepsilon$, which only appears in the weight update Eq. 7. Our averaging procedure also makes it unnecessary to scale the learning rate by the number of presentations per weight update.

**Acknowledgements**

TJS was a Wiersma Visiting Professor of Neurobiology at the California Institute of Technology when this paper was written, and we are grateful to Dr. Christof Koch for the use of his computational facilities. We especially want to thank Dr. Richard Durbin for his help in writing this syllabus and to Dr. Francis Crick for many discussions on visual processing. The work described here was supported by a Presidential Young Investigator Award to TJS, grants from the National Science Foundation, System Development Foundation, Sloan Foundation, General Electric Corporation, Allied Corporation Foundation, Richard Lounsbery Foundation, Seaver Institute, and the Air Force Office of Scientific Research.

**References**

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E., Neuronal population coding of movement direction, Science 233, 1416-1419.

Hinton, G. E., McClelland, J. L. & Rumelhart, D. E., Distributed Representations, In: Rumelhart, D. E. & McClelland, J. L., Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. Cambridge: MIT Press.

Hubel, , D. H. & Wiesel, T. N., 1962, Receptive fields, binocular interactions, and functional architecture in the cat's visual cortex, Journal of Physiology 160, 106-154.

Ikeuchi, K. & Horn, B. K. P., 1981, Numerical shape from shading and occluding boundaries, Artificial Intelligence 141-184.

Julesz, B., 1971, Foundations of Cyclopean Vision (Chicago: University of Chicago Press)

Lehky, S. R., 1983, A model of binocular brightness and binaural loudness perception in humans with general applications to nonlinear summation of sensory inputs, Biological Cybernetics 49, 89-97.

Lehky, S. R., 1987, An astable multivibrator model of binocular rivalry, (submitted for publication).

Marr, D. & Poggio, T, 1976, Cooperative computation of stereo disparity, Science 194, 283-287.

Mingolla, E. & Todd, J. T., 1986, Perception of solid shape from shading, Biological Cybernetics 53, 137-151.

O'Shea, R., 1983, Spatial and temporal aspects of binocular contour rivalry, Dissertation: Department of Physiology, University of Queensland; Brisbane, Australia.

Pentland, A. P., 1984, Local shading analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 170-187.

Rumelhart, D. E., Hinton, G. E. & Williams, R, J., 1986. Learning internal representations by error propagation, In: Rumelhart, D. E. & McClelland, J. L., Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. Cambridge: MIT Press.

Selverston, A. I., 1985, Model neural networks and behavior, (New York: Plenum Publishing).

Varela, F. & Singer, W., 1986, Neuronal dynamics in the visual cortico-thalamic pathway revealed through binocular rivalry, Experimental Brain Research (in the press).