Sejnowski, T. J. and G. E. Hinton. 1987. Separating figure from ground with a Boltzmann machine. In: Vision, Brain and Cooperative Computation, eds. M. A. Arib and A. R. Hanson, 703-724. Cambridge, MA: MIT Press.

19

Separating Figure from Ground with a Boltzmann Machine

Terrence J. Sejnowski and
Geoffrey E. Hinton

Many problems in visual processing can be formulated as searches: Given an image or a sequence of images, find the best interpretation from a large set of possible internal models. That humans are able to recognize three-dimensional objects in images within a few hundred milliseconds implies an effective search strategy. Mistakes, when they do occur, are usually confusions among similar objects. These fast, effortless, and generally reliable searches are carried out in parallel by a large number of neurons in the visual cortex. The architecture of visual cortex in primates has inspired parallel models of visual computation (Arbib 1975; Marr 1982; Feldman and Ballard 1982; Ballard et al. 1983).

In this paper we review a class of parallel visual algorithms that use relaxation to perform rapid best-fit searches and we examine some of the difficulties inherent in this search technique. In particular, we analyze the problem of separating figure from ground in an image and show how a parallel relaxation algorithm can be trapped in states that are locally optimal but globally incorrect. We introduce a general parallel search method, based on statistical mechanics, that overcomes this shortcoming and finds globally optimal solutions with a high probability (Kirkpatrick et al. 1983; Hinton and Sejnowski 1983). This approach is effective in small-scale simulations of parallel visual algorithms; its usefulness for large problems is still uncertain.

An intriguing aspect of the stochastic search procedure is that it depends on the presence of noise, which normally is considered a nuisance and which typically degrades the performance of a system. There is considerable evidence for a high degree of stochastic variability in the firing pattern of single neurons in visual cortex. This new approach raises the possibility that the noisiness of cortical neurons, rather than reflecting biological imprecision, may serve a useful purpose in improving parallel searches for optimal interpretations.

Computation of Binocular Disparity by Parallel Relaxation

Binocular depth perception, or stereopsis, has been intensively studied since Wheatstone invented the stereoscope in 1838. More recently it has been possible to study stereopsis free from other depth cues using the random-dot stereograms introduced by Julesz in 1964. Stereopsis is now known to be a difficult computational problem. Despite our much better understanding, no completely satisfactory computational solution exists, nor is there a consensus about how the problem is solved by the visual system (Mayhew and Frisby 1981; Mayhew 1983; Poggio and Poggio 1984).

Many of the issues that arise in studying stereopsis also apply to other computational problems in vision; in particular, parallel algorithms for stereopsis illustrate some of the generic difficulties of parallel visual algorithms (Ballard et al. 1983). The first step in seeing depth with two eyes is to establish matches between corresponding points on the two retinas. Matches are typically ambiguous, especially with random-dot stereograms where all local features are identical. One procedure for resolving ambiguities is to implement constraints on possible matches as excitatory and inhibitory links between processing units whose values represent depth (Sperling 1970; Julesz 1971; Dev 1975; Nelson 1975; Marr and Poggio 1976). The problem is then reduced to finding the matches that best satisfy all the local constraints.

In the Marr-Poggio (1976) algorithm for random-dot stereograms, each unit stands for a binary hypothesis about the correspondence of a particular pair of dots and therefore represents the existence of a patch of surface at a particular depth. There are exicitatory interactions between neighboring units with the same depth to ensure continuity of surfaces, and inhibitory interactions between units that represent different depths at the same image location to ensure that depth assignments are unique; if the sum of all the inputs to a unit from the two images and from local interactions is above threshold, the value of the unit is set to 1, and otherwise it is set to 0. Starting from all zeros, the units are iteratively updated: During the relaxation, various combinations of depth assignments are tried and the network eventually "locks" into a generally consistent solution in a way that resembles the human perceptual experience of fusing random-dot stereograms (Julesz 1971).

In general it is not possible to prove that this algorithm always converges to the correct depth assignments, partly because small clusters of units may form coalitions that are locally optimal but are not the globally best solution (Burt 1977; Marr et al. 1978). Another drawback of this relaxation method is the large number of iterations required to reach the final solution. If there are only nearest-neighbor interactions between units, then at least as many iterations are required as there are units across the image, since information must propagate between units one at a time and a global consensus must be reached by all the units. These problems can be minimized by the introduction of units with coarser spatial resolution (Rosenfeld and Vanderbrug 1977; Marr and Poggio 1979; Terzopoulos 1984).

Another computational problem that must be solved if only a sparse set of correct correspondences have been found is interpolating a smooth surface through the matched positions on the surface. Grimson (1981) has shown how this problem can be formulated as a variational principle in continuum mechanics by treating the surface as a thin plate. The problem is to minimize the energy of deformation of the surface constrained to pass through the matched positions. The discretized equations can be solved using a gradient-descent relaxation algorithm in which the energy is reduced at each step. As in the case of the correspondence problem, a parallel realization of the algorithm is possible with locally connected processing units. However, because in this problem the energy is a convex function possessing only a single optimum, the relaxation process always converges to the correct solution.

Special care must be taken with interpolation at locations where there are depth discontinuities. Decisions must be made either before or during the relaxation about where breaks should occur in the surface representation so that no attempt is subsequently made to interpolate smoothly across the breaks. One possibility is to monitor the local energy of deformation and "break" the thin plate if it exceeds some threshold (Terzopoulos 1984). However, once a break is made it is no longer possible to backtrack and correct for a wrong choice, so a globally optimal solution is no longer ensured. Discrete decisions must therefore be made together with the estimation of continuous variables. Similar problems occur in many other computations of intrinsic surface properties in early vision (Ballard et al. 1983).

Figure-Ground Separation

One of the simplest problems in visual perception where a discrete choice at a boundary affects subsequent processing is the organization of figure and ground in an image (Weisstein and Wong, this volume). The classic drawing that can be interpreted as either a vase or two faces (figure 1) gives

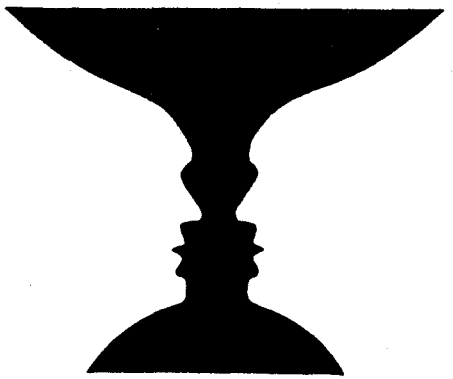

Figure 1
Rubin's (1915) demonstration of visual reversal of figure and ground. The form can be seen as a vase or as a pair of faces, but not both at the same time.

rise to two percepts depending on whether the figural part of the drawing is on the inside or the outside of the closed outline. Humans are remarkably good at performing the separation and can report within a few hundred milliseconds whether a small spot is inside or outside a briefly flashed closed outline (Ullman 1984). The discrimination probably requires two steps: a segmentation of the figure and the ground and a subsequent decision about whether the spot is located in the figure.

We briefly summarize here a simple parallel relaxation model of one type of process that occurs during figure-ground separation (Kienker et al. 1986; for previous work on scene segmentation using relaxation algorithms see Prager 1980, Zucker and Hummel 1979, and Danker and Rosenfeld 1981). The model is designed to mark the inside or the outside of a connected figure when given some lines that represent its edges and an "attentional spotlight" that provides a bias to either the inside or the outside. Examples of these two different types of input are shown in figure 3. The "bottom-up" input is not the raw image itself but is a highly processed version of the image containing the location and orientation of edges, as might be found in early visual cortex. The model must tolerate missing line segments, and it must be possible for changes in the "top-down" attentional spotlight to cause the same set of lines to be segmented differently.

There are two types of binary units in the model: figure units and edge units. Figure units correspond to small regions in the image. When a figure unit is on, its region is marked as being part of the current figure. To

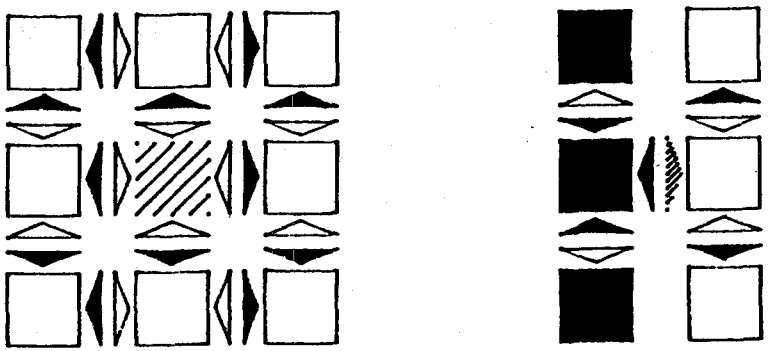

Figure 2
Summary of the weights betwen units in the figure-ground network. Because the pattern of connectivity is isotropic, only the weights for a single figure unit (left) and a single edge unit (right) are shown. The connections are represented not by conventional lines but by the presence and shading of other units. A white (open) unit represents an excitatory connection to that unit; a black (filled) unit represents an inhibitory connection to that unit. For every connection indicated there is a reciprocal feedback connection having the same weight; that is, all the connections are symmetric. Left: All the connections to a figure unit (cross-hatched square). The figure unit is connected to each of its eight nearestneighbor figure units (squares) with weights of strength +10. All the connections between the central figure unit and the surrounding edge units (arrowheads) can be deduced from the pattern (shown at right) for a single edge unit. Right: All the connections to a single vertical edge unit (cross-hatched arrowhead). The edge unit is connected to the figure unit toward which it is pointing with an excitatory weight of +12 and to the figure unit it is pointing away from with an inhibitory weight of −12. It is also connected to laterally adjacent figure units with weights of either +10 or −10. The two types of edge units, which point away from each other, are mutually inhibitory with a weight of −15. The diagram on the left shows the overall pattern of connectivity between edge and figure units.

implement the constraint that figures tend to be connected, each figure supports all eight neighboring figure units, as shown in figure 2. To implement the top-down constraint that the figure should have a particular approximate scale and a particular approximate location, figure units receive top-down excitatory input from the attentional spotlight. An edge unit is used to mark the presence and type of an edge. A line segment between two regions can be interpreted in many ways. It could be the bounding edge of a region to one side, or the bounding edge of a region to the other side, or both if it is a crack. We ignore cracks, shadows, surface markings, and edges where two non-coplanar three-dimensional surfaces join, and allow only the two alternative bounding-edge possibilities. Between any two adjacent figure units there are two edge units corresponding

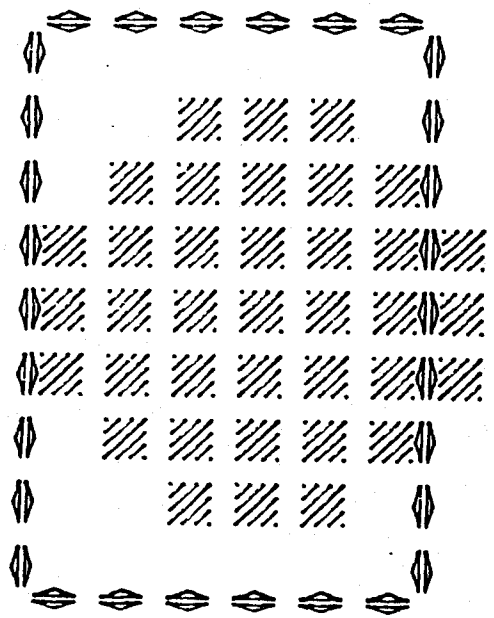

Figure 3
Two types of inputs to the figure-ground module: bottom-up inputs from the image to some of the edge units (arrowheads), which in this case form a 9 × 6 rectangle, and top-down attentional inputs to the figure units (cross-hatched squares). The strengths of the inputs to the figures units have a Gaussian distribution centered on the unit just to the right of the rectangle's center given by $15e^{-(d/2)^2}$, where d is the Euclidean distance of the unit from the center of attention. The figure units that are shown cross-hatched are those whose attentional input exceeds 1. Each figure unit has a threshold of 41, so the top-down input is not enough by itself to turn the figure units on. The edges composing the outline of the 9 × 6 rectangle have external inputs of 60, and all edge units have thresholds of 45. Thus, there was a strong bias for edge units composing the outline to be on; however, both types of edge units at each position of the outline received equal input.

to these two interpretations. Each of these supports one of the figure units and inhibits the other, and because cracks are not allowed the two edge units inhibit each other.

To implement the constraint that lines in the input require interpretation, each line segment provides equal excitatory input to the two relevant edge units. To implement the constraint that edges are implausible in places where there are no lines in the input, edge units have high thresholds that normally require excitatory input to overcome them. To implement the constraint that edges tend to be continuous, a figure unit supports the colinear neighbors of its bounding-edge units. This was found to work better than direct support between the colinear-edge units themselves, because it allows edge completion to occur around the figure region but not elsewhere.

The complete set of interactions of a figure unit and an edge unit are shown in figure 2. The precise strengths of the interactions were chosen by trial and error using a variety of outlines and were guided by the following two considerations:

- The region within the attentional spotlight should tend to be figure and the region outside should tend to be background.
- The discontinuity between figure and background should normally appear as a line in the image, and so there should be a penalty for "open frontier" where the figure region ends without there being a line in the image.

Whenever the spotlight of attention does not precisely align with the lines in the image, these two considerations are antagonistic and it therefore becomes necessary to perform a best-fit search.

One of the simplest updating algorithms consists of choosing a unit at random and summing the weighted inputs from all the active interacting units together with any external input. If this sum exceeds a fixed threshold, the unit adopts the 1 state; otherwise it adopts the 0 state. This algorithm quickly fills in the figure, but it often makes mistakes where figure units are incorrectly stabilized by edge units (as shown in figure 4). It can be made to perform reliably if the spotlight is strong; however, the performance then is very sensitive to the width of the spotlight, and this would require the top-down attentional input to already know the exact extent of the figure in the image. A more robust algorithm should be capable of good performance with a spotlight whose size and position do not already encode the exact size and position of the figure.

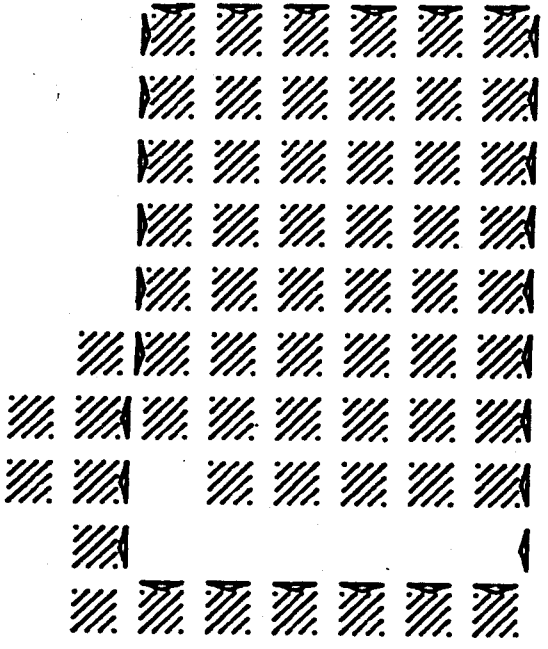

Figure 4
Final state of the figure-ground module using the gradient-descent update rule ($T = 0$). The simulation was started from a random starting state with approximately one out of ten units on. Each iteration consisted of 2,000 updates. For each update one of the 2,000 units was chosen at random, the weighted inputs from other active units were summed, and the binary threshold rule was applied to determine its new state. The system reached the steady-state configuration shown here after 28 iterations. The bottom line of figure units has been incorrectly stabilized outside the rectangle.

Analyzing Convergence

There is a useful analogy between binary networks of hypotheses that implement constraint-satisfaction problems (such as the figure-ground model introduced here) and models of interacting spins in physics. Our binary networks most closely resemble *spin glasses* (spin systems where both positive and negative interactions occur between spins). Because of competing interactions, spin glasses exhibit a phenomenon called *frustration* (Kirkpatrick 1977) in which conflicting constraints produce many local optima and degenerate ground states. One important difference, however, is that in spin glasses the spins interact randomly, whereas in binary networks that solve particular constraint-satisfaction problems the interactions are highly ordered.

The binary networks in the models of stereopsis and figure-ground separation previously discussed have the property that the connections (considered a matrix) are symmetric. A large class of constraint-satisfaction problems can be implemented with symmetric weights, including ones that require asymmetric constraints between hypotheses. For example, two hypotheses related by implication can be implemented by two units connected by symmetric weights and having different thresholds (Hinton and Sejnowski 1983). Symmetric connectivity has the significant advantage that optimization techniques and variational principles can be used to analyze the performance of the network (Hummel and Zucker 1983). In particular, Hopfield (1982) has shown that one can define an "energy" for a symmetric network of binary hypotheses that can be used to analyze its convergence. Each state is assigned an energy according to

$$E = -\tfrac{1}{2}\sum_{i \neq j} w_{ij} s_i s_j - \sum_i (\eta_i - \theta_i) s_i, \tag{1}$$

where s_i is the state of unit i, w_{ij} is the strength of connection between the units i and j, η_i is the input to unit i, and θ_i is the threshold of unit i. A simple asynchronous algorithm for finding the combination of hypotheses that has a local energy minimum is to choose asynchronously a unit at random and set its state to the one with the lowest energy. Because of the symmetric weights, this updating rule requires that the unit be set to 1 if the "energy gap"

$$\Delta E_i = \sum_j w_{ij} s_j + \eta_i - \theta_i \tag{2}$$

is positive, and to 0 otherwise. This is the familiar rule for binary threshold units that was used in describing the updating of units in the stereo

algorithm and the figure-ground algorithm. Spin models for neural networks have been studied (Cragg and Temperley 1954; Little and Shaw 1975; Choi and Huberman 1984). However, the asynchronous updating rule we have used ensures convergence, whereas the synchronous updating rule in other models may produce oscillations and more complex dynamics. Synchronous models are more closely related to cellular automata (von Neumann 1966; Wolfram 1983).

In the case of the stereo algorithm, the search space is fairly well behaved (Nishihara 1984; Prazdny 1985; Szeliski and Hinton 1985). However, the global minimum in the figure-ground problem is shallower and the search space has many local minima within which to get trapped. Many problems in vision (such as grouping and line labeling) that require the global organization of discontinuities (Waltz 1975; Zucker and Hummel 1979; Zucker 1983) have energy landscapes similar to that of the figure-ground problem. In the next section we will introduce a general technique for finding good solutions to problems of this type.

The Metropolis Algorithm and Simulated Annealing

The problem of being trapped in local energy minima can be circumvented by altering the deterministic decision rule. A simple way to escape from a local minimum is to occasionally allow jumps to states of higher energy. An algorithm with this property was introduced by Metropolis et al. (1953) for the purpose of studying the average properties of thermodynamic systems (Binder 1978). This algorithm has recently been applied to problems of constraint satisfaction (Kirkpatrick et al. 1983; Hinton and Sejnowski 1983; Smolensky 1983; Geman and Geman 1984; Bienenstock 1985). Boltzmann machines (Fahlman et al. 1983) are networks of binary processors that use as their update rule a form of the Metropolis algorithm that is suitable for parallel computation: If the energy gap between the 1 and 0 states of a unit is ΔE_i, then—regardless of the previous state—set the unit to 1 with probability

$$p_i = (1 + e^{-\Delta E_i/T})^{-1}, \tag{3}$$

where T is a parameter that acts like temperature (figure 5). As T approaches zero, equation 3 approaches a step function: the deterministic update rule for binary threshold units already introduced.

Our analysis of Boltzmann machines is based on the statistical mechanics of physical systems (Schroedinger 1946). The probabilistic decision rule in equation 3 is the same as the equilibrium probability distribution for a

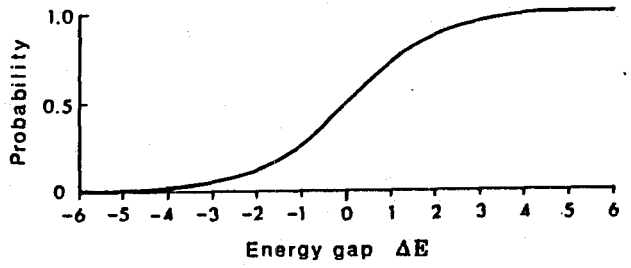

Figure 5
Probability for a unit to be on as a function of the energy gap ΔE plotted for $T = 1$ (equation 3). As the temperature decreases, the sigmoid approaches a step function; at $T = 0$ it becomes the decision rule for a binary threshold unit. As the temperature increases, the sigmoid becomes very broad and approaches a probability of 0.5 regardless of the energy gap. In this limit the effect of the weights between units becomes negligible in comparison with the thermal noise.

system with two energy states. A system of particles in contact with a heat bath at a given temperature will eventually reach thermal equilibrium, and the probabilities of finding the system in any global state will then obey a Boltzmann distribution. Similarly, a network of units obeying this decision rule will eventually reach a "thermal equilibrium" in which the relative probability of two global states of the network follows the Boltzmann distribution:

$$\frac{p_\alpha}{p_\beta} = e^{-(E_\alpha - E_\beta)/T}, \tag{4}$$

where p_α is the probability of being in the global state α and E_α is the energy of that state.

At low temperatures there is a strong bias in favor of states with low energy, but the time required to reach equilibrium may be long. At higher temperatures the bias is not so favorable but the equilibrium is reached sooner. This occurs because temperature enters as a scale factor for the energy difference in equation 4 and therefore scales the amount of discrimination between different energy states. The difficulty of breaking out of a local energy minimum depends on the heights and the degeneracies of saddle-shaped energy barriers separating them from other minima. At high temperatures these barriers are easily jumped, but lowering the temperature increases the time required to make the jump.

Kirkpatrick et al. (1983) introduced a way to find the global energy minimum using simulated annealing, a procedure derived by analogy from

the annealing of solids. The system is started at a high temperature to reach equilibrium quickly, and the temperature is gradually reduced. As the temperature is lowered, the search space is explored, first at a coarse grain and then at successively finer grains. This search procedure is effective at solving some difficult combinatorial problems such as graph partitioning, and it performs well on a large class of problems (Johnson et al. 1986). However, simulated annealing is not a panacea; there are many problems where the search space is not suitably structured. For example, it does poorly at finding a very deep and narrow energy minimum, and it would do poorly at golf (Andrew Witkin, personal communication). It is therefore not at all clear whether simulated annealing would be useful in trying to satisfy multiple weak constraints such as those found in visual algorithms.

As a test case we have applied the Metropolis algorithm and simulated annealing to the parallel algorithm for separating figure from ground introduced above. (A more detailed account can be found in Kienker et al. 1986.) At high temperatures the figure and edge units make a structureless pattern (figure 6a). As the temperature is exponentially reduced, the figure units around the center of attention tend to remain on, and these on average support those edge units whose orientation is consistent with them (figure 6b). As the temperature is further reduced, local inconsistencies are resolved and the entire network "crystallizes" to the correct solution. In a series of 1,000 annealings from random starting configurations, every trial reached the correct solution, as is shown by figure 7. A single iteration consisted of 2,000 updates in which one of the 2,000 units in the problem was chosen at random. Similar results have been obtained for a variety of simple figures, including ones where the outline is incomplete. The performance of the algorithm on spirals using the same annealing schedule is very poor; however, with a much slower annealing schedule the algorithm reliably finds the correct solution. Humans also have great difficulty with spirals.

The model of figure-ground separation presented here is clearly much too simple to explain how the problem is solved in the human visual system. A more realistic model would need to take into account multiple levels of resolution (Terzopoulos 1984) and a greater range of orientations, and it would have to introduce distinctions between low-level edge labeling and higher-level attentional phenomena (Crick 1984). However, general features of more sophisticated models are probably reflected in this simple example. More complex representations in networks of locally interacting units may also benefit from stochastic parallel search as long as

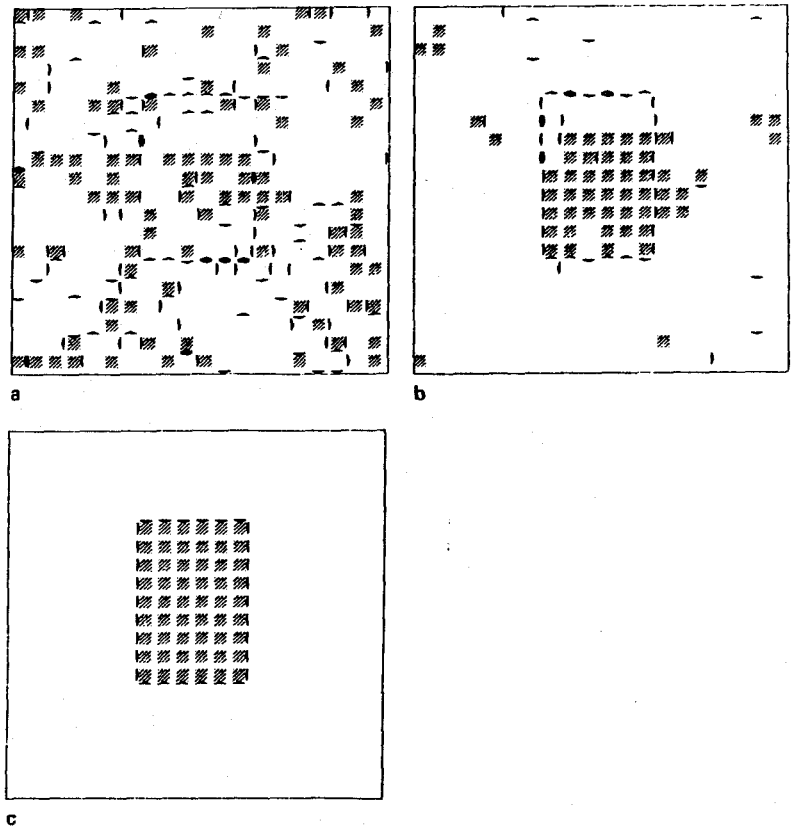

Figure 6
Simulated annealing applied to the figure-ground network shown at three temperatures—(a) $T = 16.2$ after three iterations, (b) $T = 7.7$ after ten iterations, and (c) $T = 3.3$ after 28 iterations—using the probabilistic update rule of equation 3. The annealing schedule was piecewise exponential: $T_i = p * T_{i-1}$, where $T_0 = 20$, $p = 0.9$ for $T_i > 4$, and $p = 0.99$ for $T_i < 4$.

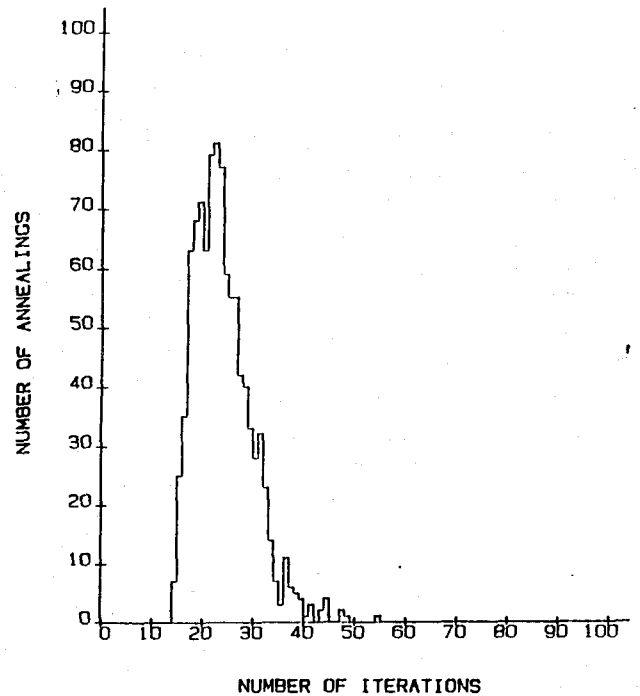

Figure 7
Histogram of the number of simulated annealing trials that properly filled the inside of the rectangle as a function of the number of iterations required. The annealing schedule given in figure 6 was used in 1,000 trials, each starting from a different random state. The fastest solution took 14 iterations and the longest took 55 iterations; the median was 21 iterations.

the best-fit solutions can be expressed as the minima of a cost function. Geman and Geman (1984) have independently used a similar approach to the Bayesian restoration of images after degradation due to blurring, nonlinear deformations, and noise.

The model of figure-ground separation used only information from the outlines of figures. Other cues, such as optical flow, may also provide information for separating figure from ground, which would require other modules. We can analyze the performance of several modules working together in parallel by simply adding together their cost functions. One of the consequences of this additivity is that different sources of evidence are weighed together linearly. It has been shown that several factors affecting the perception of depth in a rotating wire cube, including proximity luminance and perspective, are linearly additive (Dosher et al. 1985). This result is in agreement with our approach and suggests that linear additivity of evidence may be a general property of perceptual systems (Sperling et al. 1983).

Relationship between Boltzmann Machines and Neural Models

The energy gap for a binary unit has a role similar to that played by the membrane potential for a neuron; both are sums of the excitatory and inhibitory inputs, and both are used to determine the output state through a nonlinear transformation. However, a neuron produces action potentials (brief spikes that propagate down its axon) rather than a binary output. When the action potential reaches a synapse, the signal it produces in the postsynaptic neuron rises to a maximum and then decays with the time constant of the membrane (typically around 5 msec for neurons in cerebral cortex). The effect of a single spike on the postsynaptic cell body may be further broadened by electrotonic transmission down the dendrite to the spike-initiating zone near the cell body.

This suggests a neural interpretation for the binary pulses in a Boltzmann machine: If the average time between updates is identified with the average duration of a postsynaptic potential, then the binary pulse between updates can be considered an approximation to the postsynaptic potential. Although the shape of a single binary pulse differs significantly from a postsynaptic potential, the sum of a large number of pulses stochastically impinging on a processing unit is independent of the shape of the individual pulses. Thus, for networks having the large fan-ins typical of cerebral cortex (several thousand), the energy gap of a binary unit should behave like the membrane potential of a spike-producing neuron.

In addition to the nonlinear membrane currents in axons that produce action potentials, active membrane currents have also been found in the dendrites of some neurons that could support nonlinear processing (Perkel and Perkel 1985; Miller et al. 1985; Shepherd et al. 1985). This suggests that a single processing unit might be identified not with an entire neuron but with a patch of membrane. The interaction between two active membrane patches owing to electrotonic conduction on the same dendritic branch is approximately symmetrical and is always excitatory. With nonlinear interactions in dendrites, many more processing units are available; however, this advantage is partially offset by the limited topological connectivity of dendritic trees.

Noise in the Nervous System

How can the probabilistic decision rule in equation 3 be implemented by neurons? In particular, how can the temperature be controlled? The membrane potential of a neuron is graded; however, if it exceeds a fairly sharp threshold, an action potential is produced, and this potential is followed by a refractory period of several milliseconds during which another action potential cannot be elicited. If Gaussian noise is added to the membrane potential, then even if the total synaptic input is below threshold there is a finite probability that the membrane potential will reach threshold. The amplitude of the Gaussian noise will determine the width of the sigmoidal probability distribution for the neuron to fire during a short time interval, and it therefore plays the role of temperature in the model. Surprisingly, a cumulative Gaussian is a very good approximation to the required probability distribution (equation 3), never differing by more than 1 percent over the entire range of inputs.

Intracellular recordings in the central nervous system reveal stochastic variability in the membrane potentials of most neurons, which is due in part to fluctuations in the transmitter released by presynaptic terminals. Other sources of noise may also be present and could be controlled by cellular mechanisms (Verveen and Derksen 1968; Holden 1976). If some sources of noise in the central nervous system are gated or modulated, it should be possible to experimentally identify them. For example, the noise could be regularly cycled, and this would be apparent in the massed activity. Alternatively, noise may always be present at a low level and be increased irregularly whenever there is an identified need.

In the visual cortex of primates, single neurons respond to the same visual stimulus with different sequences of action potentials on each trial (Sejnowski 1981). In order to measure a repeatable response, spike trains are typically averaged over ten trials. The result, called the poststimulus time histogram, gives the probability for a spike to occur as a function of the time after the onset of the stimulus. However, this averaging procedure removes all information about the variance of the noise, so that there is no way to determine whether the noise varies systematically during the response to the stimulus or perhaps on a longer time scale while the stimulus is being attended. Such measurements of the noise variance over a range of time scales could provide evidence that this parameter has an active role in neural processing.

There are two ways to view the sigmoidal probability rule used to update units (figure 5). Over a short time interval, it represents the probability for a single unit to "fire"; over longer time intervals, in equilibrium, it represents the average "firing rate" of a unit. The average firing rate of a neuron is generally regarded as the primary neural code in the brain; however, it cannot be accurately measured over short time intervals, particularly during nonstationary conditions. The probability for a spike to occur can be defined for intervals as short as a few milliseconds and is routinely measured by ensemble-averaging spike trains, as in the poststimulus time histogram. The probabilistic interpretation of spike firing as an information code may be of more general usefulness than the average firing rate.

Symmetry, Simultaneity, and Time Delays

In a Boltzmann machine all connections are symmetrical. It is very unlikely that this assumption is strictly true of neurons in cerebral cortex. However, if the constraints of a problem are inherently symmetrical and if the network on average approximates the required symmetrical connectivity, then random asymmetries in a large network will be reflected as an increase in the Gaussian noise in each unit, in effect raising the temperature (Hopfield 1982). Systematic asymmetries that would lead to oscillations in the network would invalidate the qualitatively important feature of settling to a stable state of equilibrium.

The decision rule used in the simulations presented here was asynchronous, and updates were instantaneous. In the brain, several connected neurons may spike simultaneously within an interval of a few milliseconds. The time required for the transmission of a spike down the axon to the nerve terminal, for the release of neurotransmitter, and for postsynaptic integration can delay the signal's arrival in the spike-initiating zone of the target neuron by several milliseconds. In some simulations both simultaneous updates and transmission delays were included, and these appear to increase the noise in the system, effectively increasing the temperature (Sejnowski et al. 1985; Venkatasubramanian and Hinton 1985). At low temperatures these effects are less pronounced because the rate of flipping is lower; thus, simultaneous decisions and time delays contribute noise that could mimic annealing even without an explicit temperature control (Francis Crick, private communication). Time delays are especially effective at introducing noise, and a delay of one iteration (2,000 updates in these simultations) starting from a random state and running at $T = 1$ was almost as effective as the standard exponential annealing.

Learning in Cerebral Cortex and Boltzmann Machines

The values of weights between units for the two examples of networks discussed in this paper were chosen as much by trial and error as by plan. It would be desirable to have an automated procedure for incorporating the constraints from a given task domain in the weights. The evolution of cerebral cortex is closely linked to the ability of mammals to learn from experience and adapt to their environments; this adaptability may be the consequence of rules for modifying the strengths of cortical synapses.

A single weight between two units can be considered a "microscopic" variable in comparison with the "macroscopic" performance of the network. In general it is not possible in a network of nonlinear processing units to predict how changing a single weight will affect the overall performance. However, in a Boltzmann machine that has relaxed to equilibrium the weights between units and the probabilities of their global states are related by the Boltzmann distribution given in equation 4. Because each weight contributes independently to the energy, each weight also contributes independently in determining the relative probabilities of global states (Hinton and Sejnowski 1983). This simple probabilistic relationship makes possible a simple "microscopic" learning rule that automatically improves the "macroscopic" performance. The learning rule is similar to but different from the Hebb learning rule and has been successfully demonstrated for several small problems (Ackley et al. 1985; Hinton et al. 1984; Sejnowski et al. 1986).

Unlike the bulk of the brain, which is composed of many morphologically different nuclei, the cerebral cortex is relatively uniform in structure. Different areas of cerebral cortex (e.g. visual cortex, auditory cortex, and somatosensory cortex) are specialized for processing information from different sensory modalities, and other areas are specialized for motor functions; however, all these cortical areas are similar in anatomical organization, and they are more similar in cytoarchitecture to one another than to any other part of the brain.

The similarity between different areas of cerebral cortex suggests that massively parallel searches may also be performed in other cortical areas. Many problems in speech recognition, associative retrieval of information, and motor control can be formulated as searches. However, there is a serious obstacle that appears to prevent symmetric modules from modeling sequential information processing: At thermal equilibrium there can be no consistent sequences of states. It is tempting to use asymmetrical weights

to produce sequences, but this would be incompatible with the central idea of performing searches by settling to equilibrium.

An alternative that we are exploring is sequential settlings in a hierarchy of asymmetrically connected modules. The result of each search could be considered a single step in a strictly serial process, with each search setting up boundary conditions for the next. An attractive possibility for speeding up sequential settlings is to cascade partial settlings so that an approximate solution for one module could be used to start the search for the next one up the line (McClelland 1979). Although there are some similarities between the organization of cerebral cortex and parallel stochastic search in Boltzmann machines, more experience with larger problems and a wider range of applications are needed before the general usefulness of this approach can be properly assessed (Fahlman et al. 1983).

Acknowledgment

This research was supported by grants from the System Development Foundation and a grant from the National Science Foundation (BNS-8351331) to T.J.S. We especially thank Paul Kienker, Lee Schumacher, and Tony Yang for their assistance with the simulations.

References

Ackley, D. H., G. E. Hinton, and T. J. Sejnowski. 1985. A learning algorithm for Boltzmann Machines. *Cognitive Sciences* 9: 147–169.

Arbib, M. A. 1975. Artificial intelligence and brain theory: Unities and diversities. *Annals of Biomedical Engineering* 3: 238–274.

Ballard, D. H., G. E. Hinton, and T. J. Sejnowski. 1983. Parallel visual computation. *Nature* 306: 21–26.

Bienenstock, E. 1985. Dynamics of central nervous system. In *Proceedings of the Workshop on Dynamics of Macrosystems*, ed. J. P. Aubin and K. Sigmund (Springer-Verlag).

Binder, K. 1978. *The Monte Carlo Method in Statistical Physics.* Springer-Verlag.

Burt, P. J. 1977. A procedure for evaluating cooperative models for stereopsis. *Brain Theory Newsletter* 3: 31–34.

Choi, M. Y., and B. A. Huberman. 1984. Digital simulation of magnetic systems. *Physical Review* B 29: 2796–2798.

Cragg, B. G., and H. N. V. Temperley. 1954. The organization of neurones: A cooperative analogy. *EEG and Clinical Neurophysiology* 6: 85–92.

Crick, F. H. C. 1984. The function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences* 81: 4586–4590.

Danker, A. J., and A. Rosenfeld. 1981. Blob detection by relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3: 79–92.

Dev, P. 1975. Perception of depth surfaces in random-dot stereograms: A neural model. *International Journal of Man-Machine Studies* 7: 511–528.

Dosher, B. A., G. Sperling, and S. Wurst. 1985. Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3-D structure. *Vision Research* (in press).

Fahlman, S. E., G. E. Hinton, and T. J. Sejnowski. 1983. Massively parallel architectures for AI: NETL, THISTLE, and Boltzmann machines. In Proceedings of the National Conference on Artificial Intelligence, Washington D.C.

Feldman, J. A., and D. H. Ballard, 1982. Connectionist models and their properties. *Cognitive Science* 6: 205–254.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.

Grimson, W. E. L. 1981. *From Images to Surfaces.* MIT Press.

Hinton, G. E., and T. J. Sejnowski. 1983. Optimal perceptual inference. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, D.C.

Hinton, G. E., T. J. Sejnowski, and D. Ackley. 1984. Boltzmann Machines: Constraint-Satisfaction Networks That Learn. Carnegie-Mellon Computer Science Technical Report CMU-CS-84-119.

Holden, A. V. 1976. *Models of the Stochastic Activity of Neurones* (Lecture Notes in Biomathematics 12). Springer-Verlag.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* 79: 2554–2558.

Hummel, R. A., and S.W. Zucker. 1983. On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5: 267–287.

Johnson, D. S., C. R. Aragon, L. A. McGeoch, and C. Schevon. 1986. Optimization by simulated annealing: and experimental evaluation. In preparation.

Julesz, B. 1971. *Foundations of Cyclopean Perception.* University of Chicago Press.

Kienker, P. K., T. J. Sejnowski, G. E. Hinton, and L. E. Schumacher. 1986. Separating figure from ground with a parallel network. *Perception* (in press).

Kirkpatrick, S. 1977. Frustration and ground-state degeneracy in spin glasses. *Physical Review* B 16: 4630–4641.

Kirkpatrick, S., D. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220: 671–680.

Little, W. A., and G. L. Shaw. 1975. A statistical theory of short and long term memory. *Behav. Biol.* 14: 115–133.

Marr, D. 1982. *Vision.* Freeman.

Marr, D., and T. Poggio. 1976. Cooperative computation of stereo disparity. *Science* 194: 283–287.

Marr, D., and T. Poggio. 1979. A computational theory of human stereo vision. *Proc. Roy. Soc. Lond.* B 204: 301–328.

Marr, D., G. Palm, and T. Poggio. 1978. Analysis of a cooperative stereo algorithm. *Biological Cybernetics* 28: 223–239.

Mayhew, J. 1983. Stereopsis. In *Physical and Biological Processing of Images,* ed. O. J. Braddick and A. C. Sleigh (Springer-Verlag).

Mayhew, J. F. W., and J. P. Frisby. 1981. Psychological and computational studies towards a theory of human stereopsis. *Artificial Intelligence* 17: 349–385.

McClelland, J. 1979. On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review* 86: 287–330.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21: 1987–1092.

Miller, J., W. Rall, and J. Rinzel. 1985. Synaptic amplification by active membrane in dendritic spines. *Brain Research* 325: 325–330.

Nelson, J. I. 1975. Globality and stereoscopic fusion in binocular vision. *J. Theor. Biol.* 49: 1–88.

Nishihara, H. K. 1984. PRISM: A Practical Real-Time Imaging Stereo Matcher. Memo 780, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Perkel, D. H., and D. J. Perkel. 1985. Dendritic spines: Role of active membrane in modulating synaptic efficacy. *Brain Research* 325: 331–335.

Poggio, G. F., and T. Poggio. 1984. The analysis of stereopsis. *Ann. Rev. Neurosci.* 7: 379–412.

Prager, J. M. 1980. Extracting and labeling boundary segments in natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2: 16–27.

Prazdny, K. 1985. Detection of binocular disparities. *Biological Cybernetics* 52: 387–395.

Rosenfeld, A., and G. J. Vanderbrug. 1977. *IEEE Transactions on Systems, Man, and Cybernetics* 7: 104–107.

Rubin, E. 1915. *Synoplevede Figurer.*

Schroedinger, E. 1946. *Statistical Thermodynamics.* Cambridge University Press.

Sejnowski, T. J. 1981. Skeleton filters in the brain. In *Parallel Models of Associative Memory*, ed. G. E. Hinton and J. A. Anderson (Erlbaum).

Sejnowski, T. J., P. K. Kienker, and G. E. Hinton. 1986. Learning symmetry groups with hidden units: Beyond the perceptron. *Physica* D (in press).

Shepherd, G. M., R. K. Brayton, J. P. Miller, I. Segev, J. Rinzel, and W. Rall. 1985. Signal enhancement in distal cortical dendrites by means of interactions between active dendritic spines. *Proc. Nat. Acad. Sci.* 82: 2192–2195.

Smolensky, P. 1983. Schema selection and stochastic inference in modular environments. In Proceedings of the National Conference on Artificial Intelligence, Washington, D.C.

Sperling, G. 1970. Binocular vision: A physical and neural theory. *J. Amer. Psych.* 83: 461–534.

Sperling, G., M. Pavel, Y. Cohen, M. S. Landy, and B. Schwartz. 1983. Image processing in perception and cognition. In *Physical and Biological Processing of Images*, ed. O. J. Braddick and A. C. Sleigh (Springer-Verlag).

Szeliski, R., and G. E. Hinton. 1985. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Terzopoulos, D. 1984. Multisresolution Computation of Visible-Surface Representations. Ph.D. thesis, Massachusetts Institute of Technology.

Ullman, S. 1984. Visual routines. *Cognition* 18: 97–159.

Venkatasubramanian, V., and G. E. Hinton. 1985. On the Effects of Communication Time Delays in Boltzmann Machines. Unpublished.

Verveen, A. A., and H. E. Derksen. 1968. Fluctuation phenomenon in nerve membranes. *Proceedings of the IEEE* 56: 906–916.

von Neumann, J. 1966. *Theory of Self-Reproducing Automata*, ed. A. W. Burks (University of Illinois).

Waltz, D. 1975. Understanding line drawings of scenes with shadows. In *The Psychology of Computer Vision*, ed. P. H. Winston (McGraw-Hill).

Wolfram, S. 1983. Statistical mechanics of cellular automata. *Rev. Mod. Phys.* 55: 601–644.

Zucker, S. 1983. Computational and psychological experiments in grouping: Early orientation selection. In *Human and Machine Vision*, ed. J. Beck et al. (Academic).

Zucker, S. W., and R. A. Hummel. 1979. Toward a low-level description of dot clusters: Labeling edge, interior, and noise points. *Computer Graphics and Image Processing* 9: 213–233.

Contributors

Michael A. Arbib
University of Southern California

Dana H. Ballard
University of Rochester

Andrew G. Barto
University of Massachusetts, Amherst

Michael Brady
Oxford University

David Burr
University of Western Australia

Peter J. Burt
Radio Corporation of America
Princeton

Jerome A. Feldman
University of Rochester

Allen R. Hanson
University of Massachusetts, Amherst

Geoffrey E. Hinton
University of California, San Diego

Donald H. House
Williams College

Thea Iberall
Center for AI
Hartford

Martha Jay

Daryl Lawton
Advanced Decision Systems
Mountain View, California

Damian Lyons
Philips Laboratories
Briarcliff Manor, New York

Kenneth J. Overton
GE Corporate Research and Development Center
Schenectady

Joachim Rieger
Queen Mary College
London

Edward M. Riseman
University of Massachusetts, Amherst

David A. Robinson
Johns Hopkins Medical School

John Ross
University of Western Australia

Terrence J. Sejnowski
Johns Hopkins University

David L. Sparks
University of Alabama Medical Center
Birmingham

D. Nico Spinelli
University of Massachusetts, Amherst