Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
Washington, D. C., June, 1983

OPTIMAL PERCEPTUAL INFERENCE

Geoffrey E. Hinton

Computer Science Department
Carnegie-Mellon University

Terrence J. Sejnowski

Biophysics Department
The Johns Hopkins University

ABSTRACT

When a vision system creates an interpretation of some input data, it assigns truth values or probabilities to internal hypotheses about the world. We present a non-deterministic method for assigning truth values that avoids many of the problems encountered by existing relaxation methods. Instead of representing probabilities with real-numbers, we use a more direct encoding in which the probability associated with a hypothesis is represented by the probability that it is in one of two states, true or false. We give a particular non-deterministic operator, based on statistical mechanics, for updating the truth values of hypotheses. The operator ensures that the probability of discovering a particular combination of hypotheses is a simple function of how good that combination is. We show that there is a simple relationship between this operator and Bayesian inference, and we describe a learning rule which allows a parallel system to converge on a set of weights that optimizes its perceptual inferences.

Introduction

One way of interpreting images is to formulate hypotheses about parts or aspects of the image and then decide which of these hypotheses are likely to be correct. The probability that each hypothesis is correct is determined partly by its fit to the image and partly by its fit to other hypotheses that are taken to be correct, so the truth value of an individual hypothesis cannot be decided in isolation. One method of searching for the most plausible combination of hypotheses is to use a relaxation process in which a probability is associated with each hypothesis, and the probabilities are then iteratively modified on the basis of the fit to the image and the known relationships between hypotheses. An attractive property of relaxation methods is that they can be implemented in parallel hardware where one computational unit is used for each possible hypothesis, and the interactions between hypotheses are implemented by direct hardware connections between the units.

Many variations of the basic relaxation idea have been suggested.[1-4] However, all the current methods suffer from one or more of the following problems:

1. They converge slowly.

2. It is hard to analyse what computation is being performed by the relaxation process. For example, in some versions of relaxation there is no explicit global measure which is being optimized.

3. They are unable to integrate, in a principled way, two kinds of decision. Some systems use relaxation to make discrete decisions (e.g. which kind of 3-D edge a line depicts) and the numbers that are modified during relaxation then represent probabilities.[5] Other systems choose the most likely values of continuous physical parameters (e.g. the local surface orientation) and the numbers that are modified then represent current estimates of these parameters.[6,7] No system integrates both kinds of decision and still guarantees convergence to the optimal interpretation.

4. Systems designed to make discrete decisions do not always converge to a state in which all probabilities for discrete hypotheses are 1 or 0, so a subsequent stage is needed to choose a specific perceptual interpretation.

5. There is no obvious way for most systems to learn the appropriate values for the weighting coefficients that determine how the probabilities of related hypotheses affect each other.

In this paper we present a parallel search technique which overcomes these difficulties by using a different representation for probabilities. All the current methods use real numbers to represent the probabilities associated with hypotheses. Our method uses a more direct encoding in which probabilities are represented by probabilities. If a hypothesis has a probability of two thirds of being correct, the unit representing it will have a probability of two thirds of being found in the "true" state and a probability of one third of being in the "false" state. We first show that this direct encoding allows the probability of one hypothesis to determine the probabilities of other related hypotheses even though none of the hypothesis units ever has enough information to allow it, for example, to print out its associated probability. We then describe a search method, using this encoding, that finds plausible combinations of hypotheses. Next we show that, using our search technique, there is a Bayesian interpretation of the weights that determine the effects of one hypothesis on another, and

that the interpretation does not require the usual assumption of independence of multiple sources of evidence.

Finally we give a learning rule that allows an optimal (or near optimal) set of weights to be learnt from experience. This learning rule can be used even in cases where the representations that the system should use have not been decided in advance. The rule generates new internal representations that make explicit the higher-order statistical regularities in the environment.

Representing probabilities

There are two very different senses of the phrase "communicate a probability". In the strong sense, a unit has communicated a probability to another unit if the second unit has received enough information to allow it to print out the probability. In this strong sense, it takes a long time to communicate a probability using discrete stochastic states. To decide whether a unit is adopting the true state 100 times per second or only 90 times per second, it is necessary to observe its state for a large fraction of a second. In a tenth of a second there is only a difference of 1 in the expected number of times the unit is in the true state in the two cases. So in this strong sense, a unit that adopts truth values with a particular probability can only communicate the probability very slowly (or very inaccurately). Even if there is little physical transmission delay, there is still a long "decoding" delay before another unit has received enough information to be able to make an accurate estimate of the probability.

The decoding delay can be reduced by using a large pool of equivalent units, and by monitoring the outputs of all of them. If each unit is considered to be a Poisson process, a pool of units is a Poisson process whose rate is just the sum of the individual rates, so the decoding delay is inversely proportional to the number of units in the pool. However, the use of population averages is clearly expensive in terms of the number of units and connections required, and is therefore only worth doing if there is no more economical alternative.

Fortunately, for the kind of search we are proposing it is not necessary to communicate probabilities in the strong sense of the term. What we require is that the probability associated with unit B depends, in a particular way, on the probability associated with unit A. If these probabilites are related by some arbitrary function, it is generally necessary for unit A to communicate its probability to unit B in the strong sense of the term. But there is a special class of functions relating the probabilities of A and B that can be implemented without the units ever having to "know" (i.e. having enough information to print out) these probabilities. The simplest member of this class is the identity function. If B simply adopts the same state as A, its probability will be *exactly* the same as A's, and there will be no decoding delay. Whenever the probability associated with A changes, the probability associated with B will change after a time equal to the transmission delay alone. Another function that can be implemented this way is a probabilistic disjunction. To make the probability that unit C is in the true state be equal to the probability that either A or B is in the true state, it is sufficient to make C true if either A or B is true.

Even though the states themselves are regarded as probabilistic, the identity and disjunction functions involve a deterministic relationship between the state of one unit and the state of another. A non-deterministic relationship can be used, for example, to make the probability associated with B be half the probability associated with A. The rule is simply that B adopts the true state with a probability of one half if A is in the true state. This is a "doubly-stochastic" process in which one probability is a probabilistic function of another. We use such processes in our model of perceptual inference.

Searching for minimum energy states of a network

Given a perceptual input derived from some particular world, each possible combination of hypotheses has a particular probability of being the correct interpretation of the input. We show later that the probability can be related to a potential energy function, so that the most plausible combination of hypotheses is the one with lowest potential energy. First we give an expression for the "potential energy" of a state of a network and show how the processors have to behave in order to minimize the energy.

Hopfield[8] describes a system with a large number of binary units. The units are *symmetrically* connected, with the strength of the connection being the same in both directions. Hopfield has shown that there is an expression for the "energy" of a global state of the network, and with the right assumptions, the individual units act so as to minimize the global energy. We use a variation of Hopfield's system in which a particular task is defined by *sustained* inputs from outside the system, and the interactions between units implement constraints between hypotheses. The energy of a state can then be interpreted as the extent to which a combination of hypotheses fails to fit the input data and violates the constraints between hypotheses, so in minimizing energy the system is maximizing the extent to which a perceptual interpretation fits the data and satisfies the constraints.

The global potential energy of the system is defined as

$$E = -1/2 \sum_{ij} w_{ij} s_i s_j - \sum_i (\eta_i - \theta_i) s_i \qquad (1)$$

where η_i is the external input to the i^{th} unit, w_{ij} is the strength of connection (synaptic weight) from the j^{th} to the i^{th} unit, s_i is a boolean truth value (0 or 1), and θ_i is a threshold.

A simple algorithm for finding a combination of truth values that is a *local* minimum is to switch each hypothesis into whichever of its two states yields the lower total energy given the current states of the other hypotheses. If hardware units make their decisions asynchronously, and if transmission times are negligible, then the system always settles into a local energy minimum. Because the connections are

symmetrical, the difference between the energy of the whole system with the k^{th} hypothesis false and its energy with the k^{th} hypothesis true can be determined locally[8] by the k^{th} unit, and is just

$$\Delta E_k = \sum_i w_{ki} s_i + \eta_k - \theta_k \qquad (2)$$

Therefore, the rule for minimizing the energy contributed by a unit is to adopt the true state if its total input from the other units and from outside the system exceeds its threshold. This is the familiar rule for binary threshold units.

Using probabilistic decisions to escape from local minima

The deterministic algorithm suffers from the standard weakness of gradient descent methods: It gets stuck at *local* minima that are not globally optimal. This is an inevitable consequence of only allowing jumps to states of lower energy. If, however, jumps to higher energy states occasionally occur, it is possible to break out of local minima. An algorithm with this property was introduced by Metropolis *et. al.*[9] to study average properties of thermodynamic systems[10] and has recently been applied to problems of constraint satisfaction[11]. We adopt a form of the Metropolis algorithm that is suitable for parallel computation: If the energy gap between the true and false states of the k^{th} unit is ΔE_k then regardless of the previous state set $s_k=1$ with probability

$$p_k = \frac{1}{(1+e^{-\Delta E_k/T})} \qquad (3)$$

where T is a parameter which acts like temperature (see fig. 1).

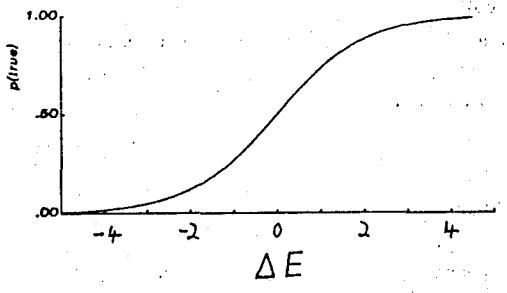


Figure 1

Probability $p(\Delta E)$ that a unit is in its "true" state as a function of its energy gap ΔE plotted for $T=1$ (Eq. 3). As the temperature is lowered to zero the sigmoid approaches a step function.

This parallel algorithm ensures that in thermal equilibrium the relative probability of two global states is determined solely by their energy difference, and follows a Boltzmann distribution.

$$\frac{P_\alpha}{P_\beta} = e^{-(E_\alpha - E_\beta)/T} \qquad (4)$$

where P_α is the probability of being in the α^{th} global state, 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 equilibrium is reached faster.

Bayesian inference

Bayesian inference suggests a general paradigm for perceptual interpretation problems. Suppose the probability associated with one unit represents the probability that a particular hypothesis, h, is correct. Suppose, also, that the "true" state of another unit is used to represent the existence of some evidence, e. Bayes theorem prescribes a way of updating the probability of the hypothesis $p(h)$ given the existence of new evidence e:

$$p(h|e) = \frac{p(h)\,p(e|h)}{p(h)\,p(e|h) + p(\bar{h})\,p(e|\bar{h})}$$

$$= 1/(1 + \frac{p(\bar{h})\,p(e|\bar{h})}{p(h)\,p(e|h)})$$

$$= 1/(1 + e^{-(ln\frac{p(h)}{p(\bar{h})} + ln\frac{p(e|h)}{p(e|\bar{h})})}) \qquad (5)$$

where $\bar{h}$ is the negation of h.

The Bayes rule has the same form as the decision rule in Eq (3) if we identify the probability of the unit with the probability of the hypothesis. The threshold implements the *a priori* likelihood ratio, the external input implements the effect of the direct evidence in the image, and the synaptic weights implement the effect of the evidence provided by the states of other hypotheses (assuming the temperature is fixed at 1):

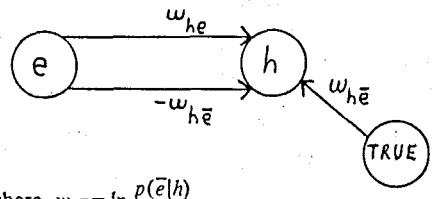
$$-\theta_h = ln\frac{p(h)}{p(\bar{h})}\ ,\quad w_{he} = ln\frac{p(e|h)}{p(e|\bar{h})}\ ,\quad \eta_h = ln\frac{p(image\,data|h)}{p(image\,data|\bar{h})}$$

Bayesian inference with one piece of evidence can therefore be implemented by units of the type we have been considering. There are, however, several difficulties with this simple formulation.

1. It provides no way for the negation of the evidence e to affect the probability of h.
2. It does not lead to symmetrical weights when two units affect each other since $p(e|h)/p(e|\bar{h})$ is generally not equal to $p(h|e)/p(h|\bar{e})$.
3. Although it can easily be generalised to cases where there are many *independent* pieces of evidence, it is much harder to generalise to cases where the pieces of evidence not independent of each other.

A diagrammatic representation of the way to solve the first difficulty is

shown below. The diagram uses a convention in which threshold terms are implemented by weights of the opposite sign on a connection from a permanently true unit. This TRUE unit is just a hypothetical device for allowing threshold terms to be treated in the same way as pairwise interactions. It simplifies the mathematics because it allows all terms in the energy expression to be treated as pairwise interactions. (The sustained external inputs that specify the particular data to be interpreted can also be turned into pairwise terms by treating them as weights on lines from units that are fixed in the true state for that particular case). The effect of $\bar{e}$ can be implemented by putting it into the threshold term for h, and by subtracting an equal amount from the weighting coefficient from e, so that when e is in the true state the effect of the threshold term on h is cancelled out.



where $w_{h\bar{e}} = ln \frac{p(\bar{e}|h)}{p(\bar{e}|\bar{h})}$

Thus the combined weight from e is:

$$w_{total} = w_{he} - w_{h\bar{e}}$$

$$= ln \frac{p(e|h)}{p(e|\bar{h})} - ln \frac{p(\bar{e}|h)}{p(\bar{e}|\bar{h})}$$

$$= ln \frac{p(e.h)[1-p(e)-p(h)+p(e.h)]}{[p(e)-p(e.h)][p(h)-p(e.h)]} \qquad (6)$$

Equation 6 is symmetrical in e and h, so in solving the problem of how to make the negation of e have the correct effect on h we have also solved the second problem -- the required weights are now symmetrical. The more complicated weight in Eq. 6 does not alter the fact that the probability of a hypothesis has the form of the Boltzmann distribution for a unit with two energy states.

Systems which use Bayesian inference often make the assumption that pieces of evidence are independent.[12,13] The main motivation for this assumption is that too much memory would be required to store all the dependencies, even if they were known. The independence assumption is hard to justify and it is typically a poor approximation in systems with many mutually interdependent hypotheses. A much better approximation, given some fixed set of variable weights, can be achieved by using whatever weights give the best overall approximation to the correct probabilities for the various possible combinations of hypotheses. At first sight, it is very hard to derive these weights, since the correct value for each weight depends on all the others. However, we now show that there are ways to hill-climb towards the optimum combination of weights.

Learning

When a system is allowed to reach thermal equilibrium using the probabilistic decision rule in Eq 3, the probability of finding it in any particular global state depends on the energy of that state (Eq 4), and so the probability can be changed by modifying the weights so as to change the energy of the state. In[14] we describe a learning rule which assumes that in addition to the input data, the system is given the desired probability ratios for pairs of global states. The rule is guaranteed to converge on a set of weights that causes the system to behave in accordance with the desired probabilities (if any such set of weights exists). We now describe a more general learning rule that does not require any separate source of information about the desired probabilities of global states. The rule leads to continual improvements in the network's model of its environment.

Suppose that the environment directly and completely determines the states of a subset of the units (called the "visible" units), but leaves the network to determine the states of the remaining, "hidden" units. The aim of the learning is to use the hidden units to create a model of the structure implicit in the ensemble of binary state vectors that the environment determines on the visible units.

We assume that each of the environmentally determined state vectors persists for long enough to allow the rest of the network to reach thermal equilibrium, and we ignore any structure that may exist in the *sequence* of environmentally determined vectors. The structure of the environment can then be specified by giving the probability distribution over all 2^v states of the v visible units. The network will be said to have a perfect model of the environment if it achieves exactly the same probability distribution over these 2^v states when it is running freely at thermal equilibrium *with no environmental input.*

In general, it will be impossible to achieve a *perfect* model because the $1/2(v+h)^2$ weights among the v visible and h hidden units are insufficient to model the 2^v probabilities of the environmentally determined states of the visible units. However, if there are regularities in the environment, and if the network uses its hidden units to capture these regularities, it may achieve a good match to the environmental probabilities.

An information theoretic measure of the discrepancy between the network's internal model and the environment is

$$G = \sum_{\alpha} P(V_\alpha) \, ln \frac{P(V_\alpha)}{P'(V_\alpha)} \qquad (7)$$

where $P(V_\alpha)$ is the probability of the α^{th} state of the visible units when their states are determined by the environment, and $P'(V_\alpha)$ is the corresponding probability when the network is running freely with no environmental input. The term $P'(V_\alpha)$ depends on the weights, and so G can be altered by changing the weights. To perform gradient descent in G, it is necessary to know the partial derivative of G with

respect to each individual weight. In most cross-coupled non-linear networks it is very hard to derive this quantity, but because of the simple relationships that hold at thermal equilibrium, the partial derivative of G is fairly simple to derive for our networks. The probabilities of global states are determined by their energies (Eq. 4) and the energies are determined by the weights (Eq. 1). Using these equations it can be shown that

$$\frac{\partial G}{\partial w_{ij}} = -\frac{1}{T}\left[\sum_{\pi} P_{\pi} s_i^{\pi} s_j^{\pi} - \sum_{\pi} P'_{\pi} s_i^{\pi} s_j^{\pi}\right]$$

where s_i^{π} is the state of the i^{th} unit in the π^{th} global state, P_{π} is the probability of the π^{th} global state (defined over both the visible and hidden units) when the network is being driven by the environment so that the states of the visible units do not depend on the weights, and P'_{π} is the probability of the π^{th} global state when the network is running freely.

To minimize G, it is therefore sufficient to increment each weight by an amount proportional to the difference between two frequencies. The first is the frequency with which the two units that the weight connects are both on when the network is being driven by the environment, and the second is the corresponding frequency when the network is running freely without environmental input. Both frequencies must be measured when the network is at thermal equilibrium. A surprising feature of this rule is that it uses only *locally available* information. The change in a weight depends only on the behaviour of the two units it connects, even though the change optimizes a global measure, and the best value for each weight depends on the values of all the other weights.

Once G has been minimized the network will be able to generate plausible completions when the environment only determines the states of *some* of the visible units. The network will have captured the best regularities in the environment and these regularities will be enforced when performing completion. One way to use this completion ability would be to divide the visible units into two subsets called "input" and "output". During "training" the environment would consist of pairs of inputs and desired outputs. In minimizing G, the network would then be finding weights that allowed it to predict the output when given the input alone.

If there are no hidden units, the weight space is concave in G so gradient descent will find the global minimum. When there are hidden units, the same learning rule still performs gradient descent in G, but there are non-global minima in the weight space, and the sytem can get stuck at one of these sub-optimal values of G. This occurs when the system is doing the best that it can given the representations it has learnt in the hidden units. To do better it has to change these representations which involves a temporary increase in G. Of course, if the modifications to the weights are probabilistic so that G can sometimes increase, it is possible to escape from local minima and ensure that after enough learning there is a bias in favor of globally optimal or near optimal sets of weights.

Potential energy and perceptual inference

In designing a parallel system for perceptual inference, the energy was important for two reasons. It represented the degree of violation of the constraints between hypotheses, and it also determined the dynamics of the search. From a few simple postulates about the energy it is possible to derive the main properties of the probabilistic system.

Postulate 1: *There is a "potential energy" function over states of the whole system which is a function, $f(P_{\alpha})$, of the probability of a state.* This is equivalent to saying that, given any input, a particular combination of hypotheses has exactly one probability. It does not, for example, have a probability of 0.3 and also a probability of 0.5.

Postulate 2: *The potential energy is additive for independent systems.* Since the probability for a combination of states of independent systems is multiplicative, it follows that $f(P_{\alpha}) + f(P_{\beta}) = f(P_{\alpha}P_{\beta})$. The only function that satisfies this equation is $f(P_{\alpha}) = k \ln(P_{\alpha})$. To make more probable states have lower energy k must be negative.

Postulate 3: *The part of the potential energy contributed by a single unit can be computed from information available to the unit.* Only potential energies symmetrical in all pairs of units have this property, since in this case a unit can "deduce" its effect on other units from their effect on it.

Discussion

We have given a brief and condensed description of a new relaxation method that overcomes many of the drawbacks of current methods. There is not space for a detailed discussion of the many interesting questions raised by the new method, and so we shall just mention a few of the more important issues here.

We have ignored the difficult question of how long it takes the system to reach equilibrium. The efficiency of the whole method depends on equilibrium being reached fairly rapidly, so this is a crucial issue. Several methods of speeding the approach to equilibrium are described briefly in[14] but more research is needed. A group at Brown University (Geman, private communication) have independently discovered the value of this kind of non-deterministic search as a model of parallel computation, and they are deriving bounds on the rate of approach to equilibrium.

It may seem disadvantageous to have a system which does not always find the most probable interpretation of the perceptual input, but instead produces interpretations with a probability that equals their probability of being correct. However, a system that integrates many different kinds of constraints will almost always pick the correct interpretation of a natural scene because with enough information the

correct interpretation is overwhelmingly more likely than any other.[15] Also, by lowering the temperature and running the system for longer it is possible to exaggerate the probability with which the most plausible interpretation will be selected.

The natural way to represent continous parameters for our relaxation method is to divide their ranges into a number of overlapping intervals and to set aside a unit for each interval[16]. The truth-value of a unit then indicates whether the continous parameter lies within its interval. By using large overlapping intervals, this representation can be made both accurate and efficient for encoding multidimensional variables.[17] An advantage of using this "mosaic" encoding is that it allows decisions about discrete and continuous variables to be integrated into a single search in a principled way.

We have ignored the fact that at finite temperature the system will inevitably settle into a "degenerate" minimum in which it fluctuates among a collection of similar states. This is actually an advantage since the proportion of the time a unit is true within the degenerate minimum allows it to convey more information about the solution than a single truth value.

We have assumed that the connections are all symmetrical in order to simplify the analysis. This assumption, however, can be relaxed. Given the symmetry of the potential energy function, it is not necessary to have two-way connections in the parallel hardware. If a symmetrical network is degraded by removing one of the directions for each pairwise link, its behavior will still approximate the behavior of the original network provided each unit has a large number of inputs, and the choice of which direction to remove for each link is random relative to the potential energy function. If these conditions hold, a unit can get a good, unbiased estimate of what its total input would have been if all the connections had been symmetrical.

A very common misconception about our relaxation method is that it is just a noisy version of continuous relaxation methods which associate a real-number with each unit. According to this view, it is the time average of the truth values that is important in the computation, and this time average can be represented by an approximate real-number. This view is wrong for several reasons. First, the computation is performed by the non-equilibrium process of reaching equilibrium, and during this process there are major differences between the ensemble average (taken over a collection of identical non-deterministic machines) and the time average (taken over time for a single machine). For example, probabilities can be accurately defined over very short time periods using ensemble averages and they can also change very rapidly. Second, the behaviour of a large ensemble of identical machines containing binary units cannot be modelled adequately by a single machine that contains real-valued units whose values represent the fraction of the corresponding units that are on in the ensemble. The single real-valued machine looses information about the higher-order statistics of the ensemble. In a case like the Necker cube, for example, there may be two alternative collections of hypotheses that form equally plausible interpretations, and a probabilistic binary machine may occasionally flip between these collections. A real-valued machine would assign a value of 0.5 to each hypothesis in either collection, and would thus fail to represent which hypothesis goes with which.

Acknowledgements

This work was supported by grants from the System Development Foundation and by earlier grants from the Sloan Foundation to Don Norman and to Jerry Feldman. We thank Dana Ballard, Francis Crick, Scott Fahlman, David Rumelhart, and Paul Smolensky, for helpful discussions.

REFERENCES

1. Davis, L. S. & Rosenfeld, A. Cooperating processes for low-level vision: A survey. *Artificial Intelligence* 1981, *17*, pp245-264.
2. Hinton, G. E. Relaxation and its role in vision. PhD Thesis, University of Edinburgh, 1977; Described in: *Computer Vision*, D. H. Ballard & C. M. Brown (Eds.) Englewood Cliffs, NJ: Prentice-Hall, 1982, pp. 408-430.
3. Hummel, R. A. & Zucker, S. W. On the foundations of relaxation labeling processes. TR-80-7, Computer Vision Lab. McGill University, July 1980.
4. Marr, D. & Poggio, T. Cooperative computation of stereo disparity. *Science*, 1976 *194*, p 283-287.
5. Faugeras, O. D. & Berthod, M. Improving consistency and reducing ambiguity in stochastic labeling: An optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1981, *PAMI-3*, pp 412-424.
6. Ikeuchi, K. & Horn, B. K. P. Numerical shape from shading and occluding boundaries. *Artificial Intelligence* 1981, *17*, pp 141-184.
7. Grimson, W. E. L. From images to surfaces. Cambridge Mass: MIT Press. 1981.
8. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA*, 1982, *79* pp 2554-2558.
9. Metropolis, N. Rosenbluth, A. W. Rosenbluth, M. N. Teller, A. H. Teller, E. *Journal of Chemical Physics*, 1953 *6*, p 1087.
10. Binder, K. (Ed.) *The Monte-Carlo Method in Statistical Physics* New York: Springer-Verlag, 1978.
11. Kirkpatrick, S. Gelatt, C. D. & Vecci, M. P. Optimization by simulated annealing. *Science* (in press)
12. Woods, W. A. Shortfall and density scoring strategies for speech understanding control. In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge Mass. August 1977, pp 18-26.
13. Yakimovsky, Y. & Feldman, J. A. A semantics-based decision theory region analyser. In Proceedings of the Third International Joint Conference on Artificial Intelligence, Menlo Park CA, 1973, pp 580-588.
14. Hinton, G. E. & Sejnowski, T. J. Analyzing Cooperative Computation. To appear in: Proceedings of the Fifth Annual Conference of the Cognitive Science Society. Rochester NY. May 1983.
15. Gibson, J. J. The perception of the visual world. Boston: Houghton Mifflin. 1950.
16. Feldman, J. A. & Ballard, D. H. Connectionist models and their properties. *Cognitive Science*, 1982, *6*. pp 205-254.
17. Hinton, G. E. Shape representation in parallel systems. In Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vol 2. Vancouver BC, Canada. 1981.