3 Holography, Associative Memory, and Inductive Generalization

David Willshaw
National Institute for Medical Research, United Kingdom

3.1. INTRODUCTION

In this chapter I review the work on the theory of associative memory that was carried out by myself in collaboration with O. P. Buneman and H. C. Longuet-Higgins at the Theoretical Psychology Unit, Edinburgh University, between 1967 and 1972. We were interested in the basic mathematical problems encountered in designing associative memory devices that would store their information in a nonlocal fashion. We were particularly interested in the design of memory models that could be implemented in neural tissue. Some of this work has already been published (Willshaw, 1972; Willshaw & Buneman, 1972; Willshaw, Buneman & Longuet-Higgins, 1969), but no overall review exists and many of the results in my thesis (Willshaw, 1971) have not been previously published.

Our task is to design structures for the storage and retrieval of items of information, which are called *patterns*. Each pattern to be stored must be identified with a second pattern, to be used as a *cue* or *address* to retrieve the first from store. We are therefore properly concerned with an *associative memory*, a device that stores *pairs* of patterns in such a way that presentation of one member of a pair will elicit the other from store. In fact the two members of a pair need not be distinct. If both were part of the same pattern, the device would be functioning as a *content-addressable* memory.

The technological advances made in the development of the hologram in the 1960s had led people to suggest that the brain functioned on holographic principles. It had long been thought that the brain might store information in a manner resistant to local damage and that allowed for correct retrieval even when an inaccurate cue was presented, and the hologram seemed to possess these prop-

erties. The key factor here is the notion of *nonlocal* or *distributed* storage: Each element of the memory contributes to the storage of more than one pair of patterns.

The logical structure of a *local* memory, such as a conventional computer store, is straightforward. Each piece of information is stored in a separate location and can be retrieved with perfect accuracy, and the quantity of information that can be stored is simply given by the number of available locations. The design of a *distributed* memory presents problems. By definition each memory location is to store a number of pattern pairs, all intermingled, and each pattern pair is distributed over more than one memory location. Thus the retrieval process involves more than just the reading out from a sequence of memory locations. One may ask, "How should information be stored in a distributed memory so that it can be retrieved accurately? Are some methods of distributed storage and retrieval better than others? How efficient is the holographic method?"

It was to answer these sorts of questions that we undertook an analysis of distributed memories. The work fell into two parts. Firstly, we investigated (Willshaw, 1971; Willshaw et al., 1969) the conditions under which distributed memories would store their information efficiently (in terms of utilization of the available memory locations) and retrieve that information with high accuracy. Holographic models were examined first, followed by other systems, some of the matrix type, which seemed to have many of the advantages of the hologram without its disadvantages. There is a simple way of classifying all the models discussed, as I explain in Section 3.5.

The second stage of the work (Willshaw, 1972; Willshaw & Buneman, 1972) examined the question of whether a memory can respond appropriately when presented with a cue to which no stored pattern corresponds. This problem can only be attempted satisfactorily if the relations between the cue and the patterns already in store are well defined. It was shown that in this case a matrix memory called the *inductive net* acts as a content-addressable memory: It supplements incomplete descriptions supplied to it on the basis of the information it has in store. Where it is logically possible for it to do so, it can supplement information with which it was not previously provided, by inductive generalization over the patterns already in store.

3.2. HOLOGRAPHIC MODELS

Holography is a method of information storage employing coherent beams of electromagnetic radiation. It was invented by Gabor (1948, 1949, 1951) and achieved technical importance with the arrival of the laser (Leith & Upatnieks, 1962; Stroke, 1966).

In this context, a *hologram* is a permanent record of the pattern of interference between two light waves, A and B, in a localized region of space. Subsequent

illumination of the hologram with one of the waves effectively unlocks the pattern from store: The incident wave, B, acts so as to cancel out the version of B previously recorded in combination with A, thus producing an approximate reconstruction of A.

Here the hologram is functioning as an associative memory. A record of A and B is stored, and information about B is used as an address to retrieve information about A. Information about A can be used similarly to retrieve information about B. Each portion of the hologram contains information about each part of A and B from which it receives light. Furthermore, if it were possible to record information about more than one pair of objects by exposing the hologram in turn to different sets of interfering waves, the hologram would be functioning as a distributed memory.

3.2.1. The Mathematics of Holography

The holographic process can be formulated mathematically by considering the following example (Collier, 1966; Stroke, 1966).

Two objects, A and B, are illuminated by monochromatic coherent light from a laser by means of a split-beam arrangement (although these remarks refer to optical systems, holograms can in principle be constructed using electromagnetic waves of any frequency). Light is reflected diffusely at the surface of the objects and interferes in a photosensitive solid whose transmittance at any point is assumed to change in direct proportion to the intensity of light at that point (Fig. 3.1). Let the complex amplitudes of the waves diffracted from objects A and B at the point x in the solid be F_A and F_B. Then the change in transmittance, Δt, at x is given by

$$\Delta t = \lambda(F_A + F_B)(F_A^* + F_B^*) \tag{3-1}$$

where λ is a numerical constant and * denotes a complex conjugate. Because F_A and F_B are complex quantities they specify the magnitude and the direction of the waves diffracted from A and B to the point x. Time variation factors have been omitted.

Object A is now removed, the photosensitive solid (the hologram) is treated so that its transmittance does not change further in response to incident light, and it is now illuminated by light reflected from object B, care being taken to maintain the spatial relationships between the parts of the apparatus. Assuming that initially the transmittance, t, is uniform throughout the hologram, the electric field amplitude, G, at x is given by

$$\begin{aligned} G &= (t + \Delta t)F_B \\ &= tF_B + \lambda\,(F_A + F_B)(F_A^* + F_B^*)F_B \\ &= tF_B + \lambda\,[(F_AF_A^* + F_BF_B^*)F_B + F_BF_A^*F_B + F_AF_BF_B^*] \end{aligned}$$

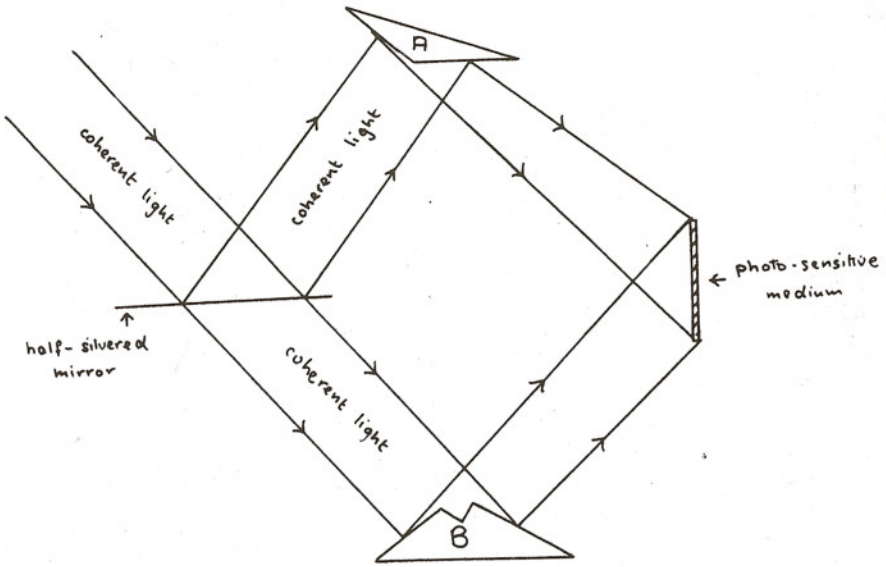

FIG. 3.1. Construction of the hologram (Willshaw, 1971).

$$\frac{G}{\lambda} = \left(I_A + I_B + \frac{t}{\lambda} \right) F_B + I_B F_A + F_B F_A^* F_B \qquad (3\text{-}2)$$

where $I_A = F_A F_A^*$ and $I_B = F_B F_B^*$ are the intensities due to objects A and B, which are assumed constant over the hologram.

This expression contains a term proportional to F_A and one proportional to F_B, which represent the wave fronts from A and B respectively. Thus, as *both* wave fronts have been reconstructed, on looking through the hologram from the right in the direction of B and then in the direction of the original position of A (Fig. 3.2) *both* images are seen. That of B is brighter than that of A, and the picture is marred by the presence of the wave fronts represented by the term $F_B F_A^* F_B$ in Eq. (3-2). As different parts of the hologram transmit these waves in different directions, the effect of this term can best be regarded as that of introducing noise into the picture.

3.2.2. Development of Holographic Models

A number of authors have drawn the analogy between holograms and the brain viewed as a distributed memory store. Van Heerden (1963a, 1963b) seems to have been the first to do so. He discussed the similarities between the method of optical information storage in solids using coherent light and Beurle's suggestions as to how associative learning could take place in a nerve net by means of

modifiable thresholds (Beurle, 1956). Van Heerden pointed out that such systems are able to store large amounts of information, and he stressed the necessity for a calibrating system of pulses in the brain in order to maintain exact phase relations between waves. Other people have discussed the importance of holography in its relationship to the experimental findings of Lashley, who had been led by them to infer that memory traces are not localized in the cerebral cortex (Lashley, 1929), and some have produced holographic brain models (Pribram, 1966; Westlake, 1968).

Longuet-Higgins extended the holographic analogy when he invented the *holophone* (Longuet-Higgins, 1968a, 1968b), which is a distributed memory working in time rather than space. In essence, the holophone is a bank of finely tuned filters, connected in parallel to a common input channel and also connected through variable gain amplifiers to a common output channel. The gains of the amplifiers make up the holophone's memory. The signal to be recorded is passed through the holophone, and the power transmitted by each of the filters is recorded. The gain of each filter is then increased in proportion to the power measured. The net result is to change the response function of the holophone so that when a small portion of a previously recorded signal is fed in, the holophone will produce an approximate rendering of the whole signal. Further signals are stored in an identical fashion.

For a detailed description of the holophone, the reader should consult the original papers of Longuet-Higgins. Its relevance to the present discussion is that it provides a simple, one-dimensional illustration of the holographic principle,

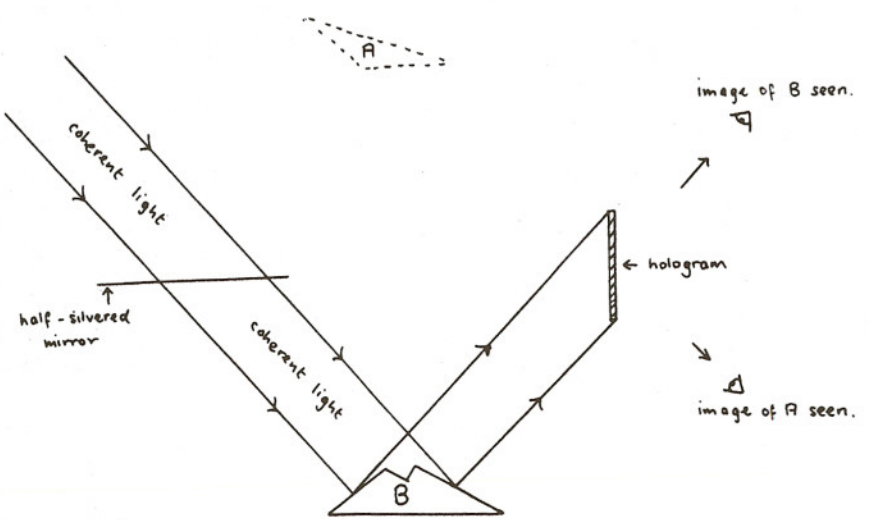

FIG. 3.2. Use of the hologram and B to produce an image of A (Willshaw, 1971).

and the mathematics derived for it are typical of other holographic memory models.

The overall computation performed by the holophone (and for that matter, other holographic devices) is relatively simple. It turns out that the effect of storing a signal in it is to change its response function by the autocorrelation function of that signal. Because, by definition of the response function, the output is the convolution of the response function with the input, the entire computation carried out by the holophone is a combination of convolutions and correlations.

The fidelity of recall was investigated by examining a discrete analog of the holophone (Willshaw and Longuet-Higgins, 1969). R patterns are stored, each represented by a sequence of N numbers. The cue is a sequence of L members of one of the stored patterns.

When each pattern value is chosen to be $+1$ or -1 with equal probability, the signal-to-noise ratio, ρ (the ratio of the square of the mean amplitude to the variance), of a component of the recalled pattern is approximately given by

$$\rho = L/NR. \tag{3-3}$$

In other words, the signal-to-noise ratio is equal to the length of the cue divided by the total length of all the patterns in store. The performance of the holophone is therefore not very good. The validity of this expression was checked by computer simulation.

This analysis of the holophone raised two questions. Firstly, holographic models of memory, requiring well-tuned filters for temporal patterns or the strict maintenance of phase relations between patterns in the spatial domain, are very complicated systems for the comparatively simple computations that they perform. Are there, therefore, simpler memory models that can mimic a holographic system? Secondly, the fidelity of retrieval of information from holographic models is very low. Are there models that can perform better? To try to answer these two questions we embarked on the work described in the next section.

3.3. NONHOLOGRAPHIC MODELS

In a search for simpler representations of holographic models of memory we took up the observation of Gabor (1968a, 1968b, 1969) that a system that computes cross-correlations or convolutions can mimic the performance of a Fourier holograph.

3.3.1. The Linear Correlograph

Fig. 3.3(a) shows such a system, called a *linear correlograph* (Willshaw, 1971). Two transparencies, A and B, are illuminated by a diffuse light source, D. Light

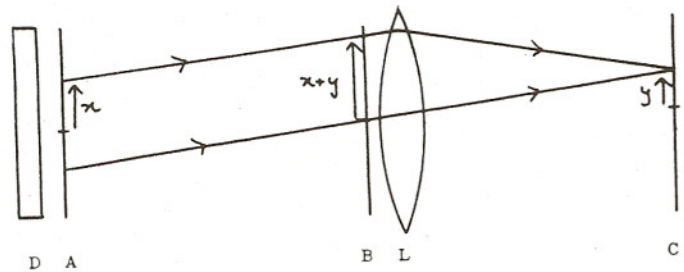

FIG. 3.3(a). Construction of the linear correlogram.

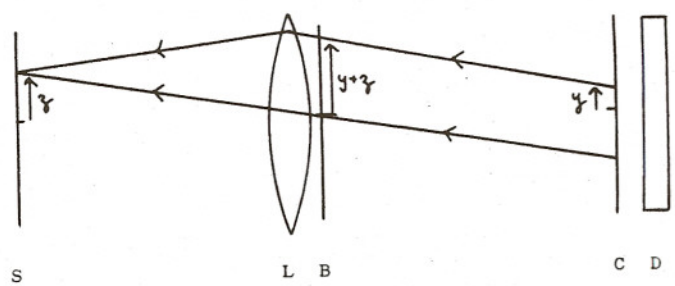

FIG. 3.3(b). Retrieval of pattern A using the linear correlogram and pattern B (Willshaw, 1971).

transmitted through B is focused onto a photographic plate C by means of lens L. A and C are separated by a distance equal to twice the focal length of the lens and B is placed halfway between them. After exposure, plate C is developed, converted into a positive transparency, which is called a linear correlogram, and then replaced. It is supposed that the transmittance at each point on plate C was modified in proportion to the intensity of the light that fell on it. The retrieval process involves illuminating the linear correlogram by the diffuse source, shifting the lens to the other side of plate B and looking at the pattern of light falling on the screen S which has replaced plate A (Fig. 3.3(b)). By recording other pairs of patterns on the same plate C, it can be made to function as a distributed memory store.

To illustrate the computations performed by this device, a simple one-dimensional case will be considered. $A(x)$ and $B(x + y)$ refer to the intensities of

light transmitted through transparencies A and B at points x and $x + y$, which contribute to the intensity, $C(y)$, at point y on C. The distances x, $x + y$, and y are measured perpendicular to the principal optic axis of the lens.

The total intensity falling at point y, and thus the amount by which the transmittance is modified at this point, is given by

$$C(y) = \int A(x)B(x + y)\, dx. \tag{3-4}$$

Retrieval also involves cross-correlations. The intensity at a point z on the viewing screen S, $S(z)$, is (disregarding numerical constants)

$$S(z) = \int C(y)B(y + z)\, dy$$

$$= \int \int A(x)B(x + y)B(y + z)\, dy\, dx. \tag{3-5}$$

Now, if R patterns $[(A^r,B^r), r = 1, 2, \ldots, R]$ are stored, the memory function C has the form

$$C(y) = \sum_{r=1}^{R} \int A^r(x)B^r(x + y)\, dx, \tag{3-6}$$

and the response to cue B^q is

$$S^q(z) = \sum_{r=1}^{R} \int \int A^r(x)B^r(x + y)B^q(y + z)\, dy\, dx. \tag{3-7}$$

This device performs the same computation as the holophone, and has therefore the same poor performance. The expression for the signal-to-noise ratio already quoted for the holophone can be readily derived if we change to discrete notation. It is supposed that each of the N components of a pattern is given equiprobably the value $+1$ or -1 and that the cue $\mathbf{b}^q$ used for the recall of $\mathbf{A}^q$ comprises L components of the vector $\mathbf{B}^q$.

The above equation can be rewritten as

$$S_{kq} = \sum_r \sum_i \sum_j A_{ir}\, B_{i+j,r}\, b_{j+k,q} \tag{3-8}$$

where the discrete quantities i, j, and k have replaced the continuous variables x, y, and z. If recall is accurate, S_{kq} should be proportional to A_{kq}.

This expression for S_{kq} is a sum of NRL products, each with value $+1$ or -1. Now L of these products each has value A_{kq} (when $r = q$ and $i = k$). Each of the other terms can, to a first approximation, be supposed to take the value $+1$ or -1

with equal probability. Thus S_{kq} has a mean value of LA_{kq} and a variance of approximately $NRL - L$. The signal-to-noise ratio is therefore

$$\rho = (LA_{kq})^2/(NRL - L),$$

or

$$\rho = L/NR \quad \text{for large } N. \tag{3-9}$$

3.3.2. The Correlograph

The linear correlograph can be radically improved by adding threshold devices to make a nonlinear system (Willshaw et al., 1969).

We now regard A and B as black cards on which are punched patterns of pinholes. These when illuminated by the light source D produce a pattern of spots on card C. Information is stored by punching a hole through C at every point where a spot appears. For example, suppose that A has 2 holes and B has 3. 6 spots will appear on C, some of which may coincide (Fig. 3.4).

On reconstruction of A, 18 rays strike S; 6 of these retrace the path of rays that originally passed from A through B onto C, and the other 12 strike S at spurious points, points that were not originally pinholes in card A (Fig. 3.5). Consequently, pattern A is seen amongst a background of noise. However because genuine points on S (those which correspond to pinholes on card A) receive 3 rays while spurious points receive fewer than 3, a threshold detector set at a value of 3 will produce a faithful reproduction of pattern A.

Other pairs of patterns are then added by punching holes in card C at the relevant points. In fact, some of these holes will have already been made in the storage of previous pairs. Provided that the number of stored pattern pairs is kept below a certain level, retrieval is excellent, as I now demonstrate.

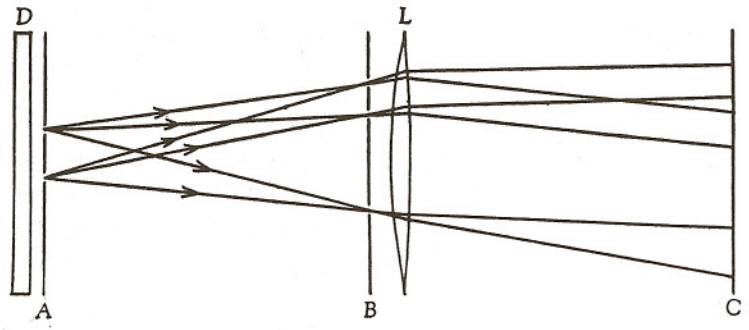

FIG. 3.4. Constructing a correlogram. D is a diffuse light source, L is a lens, and C is the plane of the correlogram of A with B (Willshaw, et al., 1969).

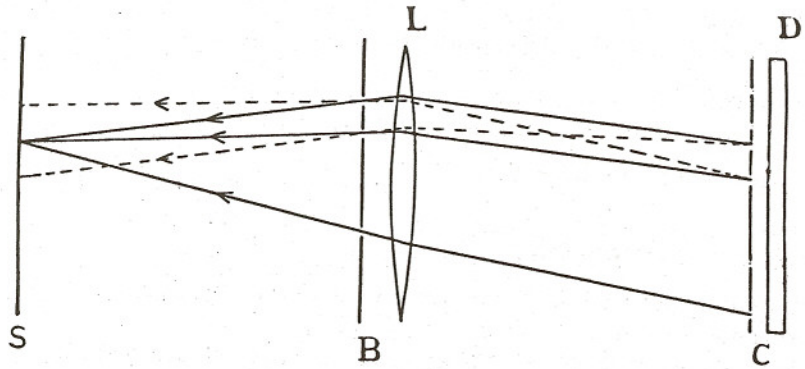

FIG. 3.5. Reconstructing a pattern. Full lines are paths traversed in Fig. 3.4; broken lines are paths not traversed. There are 18 rays that travel from C to S; 5 of these are shown here (Willshaw et al., 1969).

To investigate the conditions for good recall, we adopt an abstract discrete representation. A, B, and C are discrete spaces, each of N points. The N^2 point-pairs formed by linking one point from A with one from B are mapped onto C in the following manner. Point-pair (a_i, b_j) maps onto c_k if $k = (j - i)$ or $(j - i + N)$. Thus the N^2 point-pairs are distributed equally amongst the N points of C. Conversely, in the retrieval process (c_k, b_j) is mapped onto a_i if the same condition is met.

There are R pairs of patterns, each comprising a selection of M of the members of A and M of the members of B. There are therefore RM^2 pairs of points spread out at random amongst the N points of C. The probability, p, that a point c_k is identified with at least one of these pairs is given by

$$1 - p = e^{-RM^2/N} \tag{3-10}$$

The number of points in C identified with the R pattern pairs is therefore pN.

In the reconstruction process, one B-pattern, comprising M points, is combined with the correlogram C to produce a number of rays striking S. Each genuine point on S will receive M rays, so that a detector set to respond to M or more rays will sort out these points from the noise. There is, however, a chance that other positions on S will receive M rays and so register a false report. This will happen with probability p^M, so that the mean number of spurious points in the retrieved pattern will be

$$(N - M)p^M.$$

Good retrieval will be ensured if this number is no greater than 1. A slightly safer upper limit is

$$Np^M = 1,$$

or

$$M = -\log N/\log p. \tag{3-11}$$

We are now in a position to determine the conditions under which the system works optimally. In the retrieval of R patterns, each made up of a selection of M out of N points, the information gained is

$$I = R \log_2 \binom{N}{M} \text{ bits}$$

$$\simeq RM \log_2 N \text{ bits}. \tag{3-12}$$

Using Eq. (3-10) and (3-11), which are expressions for R and M, it follows that

$$I = N \log_2 p \log_e (1-p) \text{ bits}, \tag{3-13}$$

which has a maximum at $p = .5$,

$$I/N = \log_e 2 \text{ or } 0.693$$

and

$$M = \log_2 N. \tag{3-14}$$

Therefore this device works most efficiently when the number of points M in a pattern is related logarithmically to the number of possible points N and when half of the N points of the correlograph have been converted into pinholes. Furthermore, because the correlogram C is effectively a binary store of N bits, under these conditions this device works 69% as efficiently as a conventional store with no associative capability.

3.4. THE ASSOCIATIVE NET

In the previous section I described how the essentials of the holographic memory can be distilled into a simpler model called the correlograph. Its information is stored in a distributed fashion, and with the addition of threshold detectors it can be used almost as efficiently as a conventional local store.

The correlograph has another property, however, which poses a difficulty when one considers how it could be represented in the nervous system. This is the fact that it can cope with displaced patterns. In the reconstruction mode, if the

pattern B to be used as a cue is a spatially displaced version of one previously stored, the output will be the appropriate A-pattern, also displaced. The reason for this is that in the construction of the correlogram there is a many-to-one mapping of the point pairs (A, B) onto C: Each point C is identified with N of the N^2 point pairs made by combining members of A with members of B.

This facility is absent in a memory system where each storage location is identified with a unique point pair. Such a system, called an *associative net* (Willshaw, 1971; Willshaw et al., 1969) can be represented by a lattice (Fig. 3.6). The N_A vertical lines and the N_B horizontal lines represent the N_A points of A and the N_B points of B, and the $N_A N_B$ intersections represent the points of C. A particular point in C is regarded as being active if the pair of lines (a_i, b_j) that pass through it have been called into play in the association of at least one of the R pattern pairs. The mathematics of this device is similar to that of the correlograph. Let us suppose that each pair is a selection of M_A of the N_A lines and M_B of the N_B lines. The probability, p, that a given point of C has been activated in the recording process is given by

$$1 - p = e^{(-RM_A M_B / N_C)} \tag{3-15}$$

where N_C is written for $N_A N_B$.

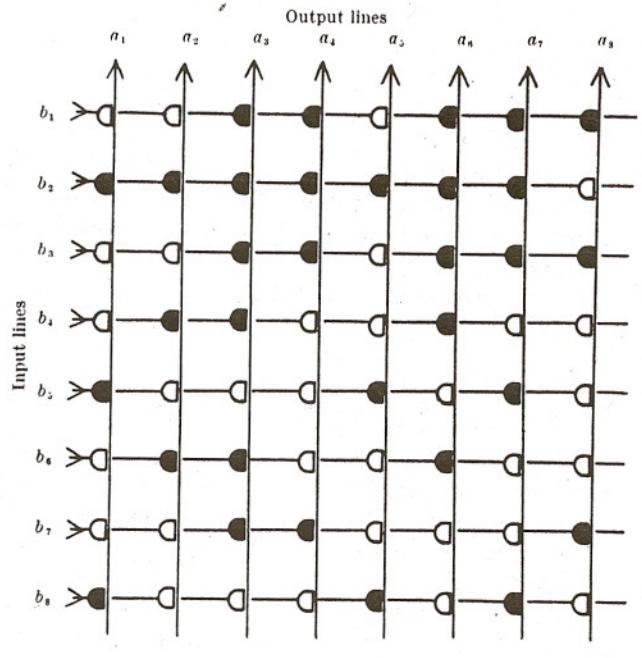

FIG. 3.6. An associative net. The nodes that have been activated in the storage process are colored black (Willshaw et al., 1969).

When B-patterns are used to recall A-patterns, the threshold detector must be set at M_B units. It will fire in response to spurious outputs with probability p^{M_B}, so that the limit of good recall is set at

$$N_A p^{M_B} = 1. \tag{3-16}$$

It follows that the amount of information stored in the memory is

$$I_A = N_C \log_2 p \log_e (1-p) \text{ bits.} \tag{3-17}$$

As in the correlograph, the efficiency with which the associative net can be made to store information is 69% of the theoretical maximum. I have analyzed the performance of the associative net under a variety of conditions. In some cases I carried out computer simulations to check the analysis. Some of the results are described below; for other results the reader should consult Willshaw (1971).

3.4.1. The Associative Net as a Neural Model

There is a straightforward way of realizing the associative net in nervous tissue. The horizontal lines of Fig. 3.6 are axons of the N_B input neurons, b_1, b_2, . . ., and the vertical lines are dendrites of the N_A output neurons, a_1, a_2. . . . At the intersection of a_i with b_j is a modifiable synapse c_{ij}. This synapse is initially inactive but becomes active after a_i and b_j are made to fire simultaneously, which will occur if an A-pattern containing a_i is presented in association with a B-pattern containing b_j. Thus the synapse functions as a binary switch. If c_{ij} has been activated during the storage procedure, the firing of b_j in the retrieval mode will locally depolarize the membrane of a_i. The output neuron a_i is then supposed to fire if M_B or more input cells depolarize it simultaneously.

The regular network structure of the nerve cells of the cerebellar cortex (Eccles, Ito, & Szentágothai, 1967) might suggest that this part of the nervous system is a form of associative net. But there are important differences between the cerebellum and the associative net that would reduce the efficiency with which the cerebellum could be used in this manner. In particular, the anatomy of the cerebellum dictates that the threshold on the firing of the output lines (Purkinje cells) cannot be set exactly equal to the number of actual input lines but must be determined by a sampling procedure. For a detailed proposal of how the cerebellum might be used as a learning device the reader is referred to the papers by Marr (1969) and Blomfield & Marr (1970). To my knowledge their key proposal, that synapses between the parallel fibers and the Purkinje cells are modifiable, has yet to be confirmed.

One property of the associative net that makes it attractive as a neural model is that good retrieval can be obtained even when some of the storage elements are damaged or when some of the components of the address are incorrect. To achieve this the number of lines to be activated in the storage of a pattern must be

greater (and the loading of the net must be lighter) than under optimal conditions (Willshaw, 1971). Resistance to local damage is bought at the expense of storage capacity.

3.4.2. The Associative Net with a Feedback Loop

The nonlinearities of the associative net give it interesting properties when it is equipped with a feedback loop. Here, in the retrieval mode the response to a given cue is used as a new input. This is well illustrated when the associative net is used as a content-addressable memory. Patterns of the type (A, A) are stored and a fragment A' is used to retrieve the rest of A from store. It was found by computer simulation (Willshaw, 1971) that the initial response to a given cue could be improved by feeding the output back into the associative net and continuing until the sequence of outputs so generated converged onto a single pattern.

The same "cleaning-up behavior" was seen when patterns were stored in sequence. Pattern A was associated with B, B with C, C with D, and so on, the last pattern being stored with A. When a fragment of A was used as a cue and then the output used as the next input, after a few passes the sequence of retrieved patterns converged onto the stored sequence, even when the initial cue was a very poor representation of one of the stored patterns. Simulation experiments were performed to see what cycle of outputs would result from any arbitrarily selected cue. (Because each input determines the next output and there is only a finite number of possible outputs, the sequence of outputs must eventually lead into a cycle.) It turned out (Willshaw, 1971) that the length of the typical cycle is very small. For an associative net with 64 horizontal and 64 vertical lines, the length of the cycle generated from an arbitrarily chosen input was typically 50. This is of the order of magnitude of the logarithm of the number of different possible output patterns rather than that number itself, which is 2^{64} or approximately 10^{20}. In all cases when the associative net is equipped with a feedback loop, good recall can only be achieved by loading it more lightly than the optimal conditions, derived in Section 3.4, permit.

3.5. A COMPARISON OF CORRELOGRAPHIC AND MATRIX MODELS

Two sorts of distributed memory have been described. The one has a bank of N memory elements to store associations between the points of A and the points of B, each of which has N elements; the other has N^2 memory elements. What is the formal relationship between these two memory systems, and do they represent the main types of distributed memory?

Let us return to our starting point and consider a way of representing the holographic method of storage in matrix notation. The N-dimensional vector

$\mathbf{A}$ specifies the complex amplitude distribution of the light wave, sampled at N points, which is diffracted from object A. In the holographic plane this amplitude distribution becomes

$$\boldsymbol{\Psi} = \mathbf{MA} \tag{3-18}$$

where $\mathbf{M}$ is the matrix of the Fourier transformation.

A similar equation can be written for the wave diffracted from object B.

$$\boldsymbol{\Phi} = \mathbf{MB} \tag{3-19}$$

The construction of the hologram involves increasing the transmittance of each of the N points of the holographic plate in proportion to the intensity of light resulting from the interference of the two waves, which is a function of the two vectors $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$. The changes in transmittance in all the points of the holographic plate can be expressed in terms of the $N \times N$ storage matrix $\mathbf{X}$, which in this case is a diagonal matrix. Subsequent storage of other pairs of patterns leads to further changes in transmittance, and therefore to further alterations to the matrix $\mathbf{X}$.

We can now formulate an equation for the whole process. Given that a number of pairs of patterns have been stored, suppose that pattern $\mathbf{B}$ is used as the cue to recall $\mathbf{A}$. On presentation of $\mathbf{B}$, the amplitude distribution $\boldsymbol{\Phi} = \mathbf{MB}$ is set up at the hologram. This distribution is then modified by the memory matrix $\mathbf{X}$. Finally, on moving back from Fourier space into object space, the inverse Fourier transform, described by the matrix $\mathbf{M}^{-1}$, is performed. These three steps are described by the equation

$$\mathbf{S} = \mathbf{M}^{-1}\mathbf{XMB} \tag{3-20}$$

This equation describes in the most general terms the class of *linear* nonlocal associative memories. Different devices can be constructed by choosing different nonlocal transformations, $\mathbf{M}$, and different forms for the storage matrix, $\mathbf{X}$. The corresponding class of *nonlinear* associative memories is described by the equation

$$\mathbf{S} = [[\mathbf{M}^{-1}\,\mathbf{X}\,\mathbf{M}]\mathbf{B}] \tag{3-21}$$

where the square parentheses [] indicate nonlinear operations.

Two special cases of these equations are now considered.

As already indicated, substituting the holographic forms of $\mathbf{M}$ and $\mathbf{X}$ in equations 3-20 and 3-21 leads to the mathematics of the holophone and the correlograph respectively. In fact, matrix $\mathbf{M}$ need not be the Fourier matrix; using any orthogonal matrix will lead to the same result.

Secondly, let us suppose that the holographic method of calculating the components of the storage matrix $\mathbf{X}$ is applied to each of its components rather than to the diagonal ones only. $\mathbf{X}$ will therefore have N^2 rather than N nonzero components. Then it is straightforward to calculate (Willshaw, 1971) that the response to the cue $\mathbf{B}^q$ is

$$S_{iq} = \sum_r \sum_j A_{ir} B_{jr} B_{jq}. \tag{3-22}$$

This is the correlation type of memory, which has also been discussed by Anderson and by Kohonen. In this model, the signal-to-noise ratio of the response to a cue of length L from a memory that has stored R pattern pairs each of length N is

$$\rho = L/R, \tag{3-23}$$

which is a factor of N greater than the figure for the holophone. Once again, by inclusion of nonlinear operations, this model can be converted into the associative net. Here, too, the form of the matrix, **M**, is not crucial.

3.6. THE INDUCTIVE NET

The associative net is able to deal with incomplete and inaccurate information. When a distorted version of one of its stored patterns is presented to it, under certain conditions the correct pattern can be retrieved accurately. We now take the argument a stage further. The question is whether a memory can be designed, which, when given an incomplete description of a stored item, will furnish those details missing from the description; and will also accept a description to which no stored item corresponds and will supplement this description by inductive generalization over the items already in store. This question is meaningless unless the relationships between the items of information to be stored are well defined. Suppose I am given some numbers to memorize. I am then given another number with a few missing digits and am asked to fill in the gaps on the basis of the order in the numbers I have already learnt. I shall only be able to do this if I can infer the rules used in constructing the given numbers; if there are no rules, the task is insoluble.

The ensemble of patterns that we consider is made up of binary vectors of fixed length. Each vector contains a fixed number of binary features, and for each pair of features not all of the four combinations of feature values, "++", "+−", "−+", "−−", occur amongst the ensemble of vectors. Suppose, for example, that the vectors describe the appearance of wooden blocks and that one feature relates to color, blue or green, and the other to size, large or small. Further, only three of the four combinations are allowed, there being no blocks which are both green and small. Then if we are told that a particular block is small, we can infer that it is blue; and if a block is green, it must be large. Information about one feature can be used to infer information about another. No two features are logically independent of each other; such an ensemble is said to obey the *four-point condition* (Buneman, 1971). It turns out that the members of

such an ensemble can be represented by an unrooted tree, each of whose nodes represents a vector. Each link represents a feature, and so separates those vectors with value "+" for this feature from those with value "−". A six-link tree is shown in Fig. 3.7, and this defines an ensemble of seven binary vectors.

When a selection of patterns is made from a four-point ensemble, owing to the logical relations between the various features more information about the ensemble can be inferred than was explicitly given. For example, let us choose from Fig. 3.7 the patterns **A, E, F,** and **G,** which are

A: (+1, −2, +3, −4, −5, +6),

E: (+1, −2, −3, +4, +5, −6),

F: (+1, +2, +3, +4, +5, −6),

G: (−1, −2, +3, −4, +5, −6).

It can be shown that there are two four-point ensembles from which these four patterns could have been drawn. Each ensemble can be represented by a tree; the two ensembles are represented by the superposition of the two trees, called a *multitree* (Fig. 3.8), where the relative positions of links five and six are not determined. Given that the patterns are from a four-point ensemble, the existence of patterns not explicitly stored can be inferred, as shown by the presence of the nodes labeled **C** and **D** in Fig. 3.8.

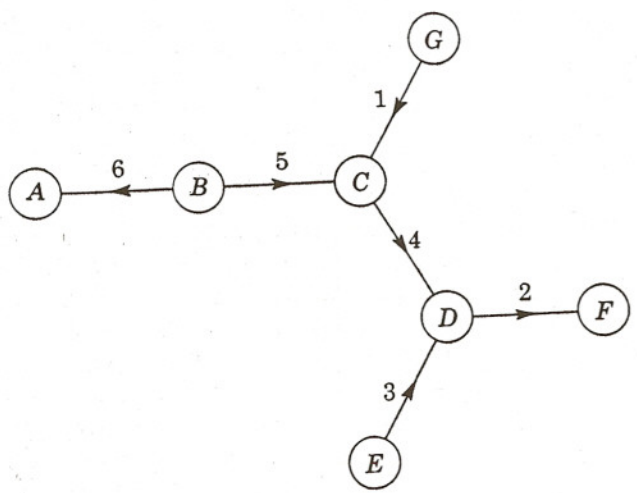

FIG. 3.7. A six-link binary tree. As an illustration of how vectors are assigned to the nodes of a binary tree, the six-dimensional vectors identified with **A** and **D** are: **A**: (+1, −2, +3, −4, −5, +6) **D**: (+1, −2, +3, +4, +5, −6) (Willshaw, 1972).

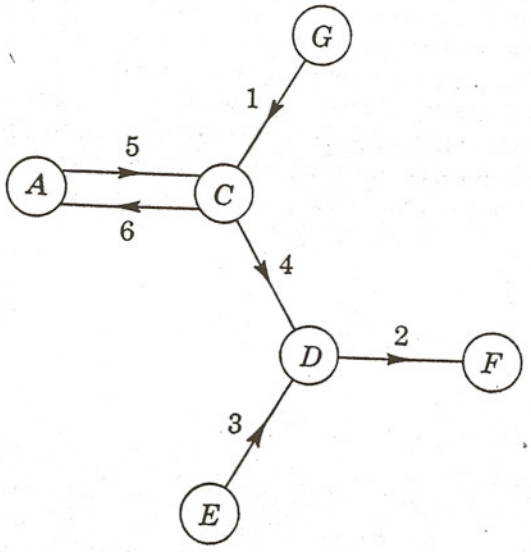

FIG. 3.8. The multitree constructed from the vectors identified with nodes **A, E, F,** and **G** of Fig. 3.7 (Willshaw, 1972).

The network that can perform inductive generalizations of this kind is similar to the associative net and is shown in Fig. 3.9. This *inductive net* (Willshaw, 1972) is a content-addressable memory, and it is similar in structure and function to the associative net. In the storage of a pattern, the memory elements that are in the inactive state and lie at the intersections of the active horizontal and vertical lines are switched on. In retrieval, activity in the horizontal lines associated with the cue stimulates, through the switches previously turned on, certain (vertical) output lines, and a threshold detector set at the number of active horizontal lines causes some output lines to fire.

In this network there are two horizontal lines associated with each feature, one with each feature value. The vertical lines are also arranged in pairs. Activity in the two vertical lines associated with the same feature is mutually inhibitory, so that there can never be responses from both lines simultaneously.

Let us now see how the inductive net functions. The inductive net of Fig. 3.9 has stored the descriptions of the vectors **A, E, F,** and **G.** We now present an incomplete description of one of the vectors, say **F,** by activating input line +4. The output is the set of values (+1, +4, +5, −6). Reference to the multitree (Fig. 3.8) shows that this set of feature values is indeed the set common to those vectors that have the value +4. This retrieved set does not specify vector **F** uniquely because there are other vectors that have value +4. The inductive net can therefore supplement incomplete descriptions of stored items, as far as it is logically possible for it to do so.

A more illuminating example is provided by using the cue (−2, +3, +4), which specifies the vector **D.** The output is (+1, −2, +3, +4, +5, −6), a complete description of **D.** What is interesting is that **D** was not explicitly stored. This illustrates the point that the inductive net can generalize, where logically possible, to complete descriptions of items not explicitly stored. Proofs of the theorems specifying the performance of the inductive net can be found in Willshaw (1972).

The capability of the inductive net to generalize becomes an embarrassment when the patterns whose presence are inferred are not in fact members of the chosen ensemble. The general question here is: How does the inductive net deal with ensembles that do not obey the four-point condition? Peter Buneman and I (Willshaw & Buneman, 1972) have considered how to adapt this device to solve the *exact match* problem, which is that of searching amongst a set of stored patterns to find those specified by a given partial description. The system is required to respond only on the basis of the information explicitly stored; there is to be no generalization. We were interested in a content-addressable memory that would work perfectly rather than in the statistical manner of models such as the associative net.

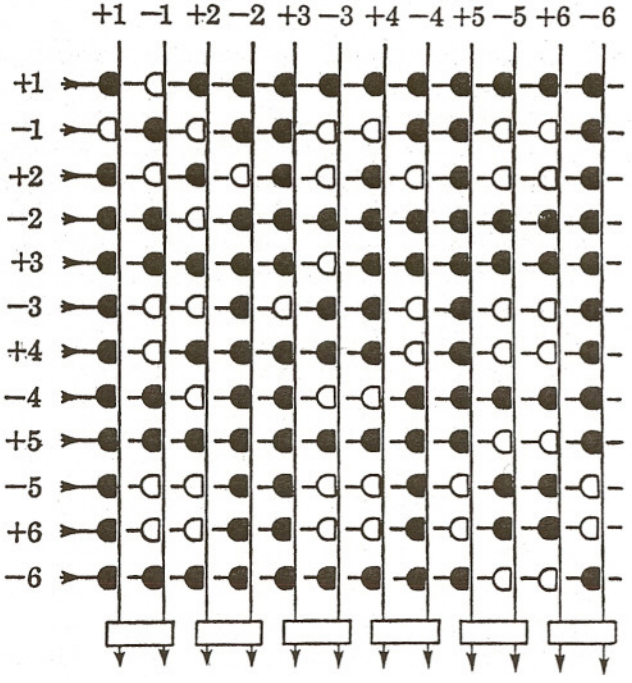

FIG. 3.9. The inductive net associated with Fig. 3.8 (Willshaw, 1972).

An inductive net can be adapted to deal with ensembles violating the four-point condition by adding extra input lines that respond to *combinations* of feature values; we talk about adding a series of *masks* of various *sizes*. The size of a mask is the number of feature values which must be looked at in order to decide whether the input line should fire. We have shown that the information about the vectors to be stored can be represented on a type of graph that is a generalization of the multitree and that an inductive net equipped with the appropriate extra input lines will be able to supplement incomplete descriptions of the patterns in store. An as-yet-unsolved problem is to find an efficient way of selecting the extra masks. Unless there are very few features it would be inefficient to include all possible masks because the number of possible masks is exponentially related to the number of different features.

3.7. CONCLUSION

It is possible to design nonlocal memory devices that store and retrieve their information efficiently. In considering the various different types of memory, useful distinctions can be made: (1) between the correlographic and the matrix types; and (2) between linear and nonlinear systems, as follows.

1. Matrix versus correlographic memories. In a matrix memory, each storage location is responsible for a different point-pair. Thus pairs of patterns each with N components are stored in a memory with N^2 registers. This system will be therefore expected to have a storage capacity N times that of the correlographic type, which has just N registers. This is certainly true when one compares the associative net with the correlograph. A similar relation exists between the linear versions of the two systems. In both cases the signal-to-noise ratio varies inversely with the number of stored patterns, and for the linear associative net the number of patterns needed to reach a given signal-to-noise ratio is N times that needed for the linear correlograph.

2. Linear versus nonlinear memories. Linear memory models can be described by the matrix equation

$$\mathbf{S} = \mathbf{M}^{-1}\mathbf{XMB} \tag{3-20}$$

and nonlinear models by the equation

$$\mathbf{S} = [[\mathbf{M}^{-1}\mathbf{XM}]\mathbf{B}]. \tag{3-21}$$

A particular memory model is defined by a particular choice of $\mathbf{M}$, the transformation matrix, and $\mathbf{X}$, the memory matrix. There is nothing remarkable about the holographic type of model. Indeed it seems to be overcomplex for the computations it performs.

Linear models appear to have disadvantages when used for distributed storage. By its very nature, distributed storage intermingles the stored patterns, and a linear model's response to a given cue will be an average of all stored patterns; even with very few patterns, a retrieved pattern may resemble none of the patterns in store. Linear memories can only be made to work reliably when there are heavy constraints put on the ensemble of patterns to be stored, such as that the patterns form an orthogonal set of vectors.

Nonlinear systems, such as the associative net, can be made to function almost perfectly as long as the memory is not heavily loaded. To use these systems efficiently, however, constraints must be imposed on the type of pattern that can be stored. In the case of the associative net, the number of lines active in storing a pattern must be logarithmically related to the total number of lines available. This constraint is not as severe as those required for the linear model, but it is a constraint nevertheless.

What are the problems in applying the theory of associative memory, either to the design of commercial devices or to the analysis of nervous function? As I see them, the problems are not concerned with the logic of the distributed memory devices as such, but with how to encode the patterns to be stored into a form acceptable to the memory and how to convert the output of the memory back into one of the given patterns. A general-purpose encoding and decoding algorithm, such as one that orthogonalizes the patterns to make them suitable for a linear system, would be very difficult to devise. As far as biological applications are concerned, instead of treating biological patterns as strings of random digits, it would be worth investigating their structure, that is, the logical relations between their component parts. As our work on the inductive net has shown, if the logical relations underlying the data can be isolated, memory systems can be designed that exploit this structure and thereby acquire properties reaching beyond those of a simple memorizing device.

ACKNOWLEDGMENTS

I thank MacMillan Journals Ltd. for permission to reproduce Figs. 3.4, 3.5, and 3.6 and the Royal Society of London for permission to reproduce Figs. 3.7, 3.8, and 3.9.

REFERENCES

Beurle, R. L. Properties of a mass of cells capable of regenerating pulses. *Philosophical Transactions of the Royal Society. Series B* 1956, *240*, 55-94.

Blomfield, S. & Marr, D. How the cerebellum may be used. *Nature*, 1970, *227*, 1224-1228.

Buneman, O. P. The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G.

Kendall, & P. Tautu (Eds.), *Mathematics in the archaeological and historical sciences*. Edinburgh University Press, 1971.

Collier, R. J. Some current views on holography. *I.E.E.E. Spectrum*, 1966, *3*, 67–74.

Eccles, J. C., Ito, M., & Szentágothai, J. *The cerebellum as a neuronal machine*. Berlin: Springer-Verlag, 1967.

Gabor, D. A new microscopic principle. *Nature*, 1948, *161*, 777–778.

Gabor, D. Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society. Series A* 1949, *197*, 454–487.

Gabor, D. Microscopy by reconstructed wavefronts. II. *Proceedings of the Physics Society*, 1951, *64*, 244–255.

Gabor, D. Holographic model of temporal recall. *Nature*, 1968, *217*, 584. (a)

Gabor, D. Improved holographic model of temporal recall. *Nature*, 1968, *217*, 1288–1289. (b)

Gabor, D. Associative holographic memories. *IBM Journal of Research and Development*, 1969, *13*, 156–159.

van Heerden, P. J. A new optical method of storing and retrieving information. *Applied Optics*, 1963, *2*, 387–392. (a)

van Heerden, P. J. Theory of optical information storage in solids. *Applied Optics*, 1963, *2*, 393–400. (b)

Lashley, K. S. *Brain mechanisms and intelligence*. Chicago: University of Chicago Press, 1929.

Leith, E. N. & Upatnieks, J. Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America*, 1962, *52*, 1123–1130.

Longuet-Higgins, H. C. Holographic model of temporal recall. *Nature*, 1968, *217*, 104. (a)

Longuet-Higgins, H. C. The non-local storage of temporal information. *Proceedings of the Royal Society. Series B* 1968, *171*, 327–334. (b)

Marr, D. A theory of cerebellar cortex. *Journal of Physiology*, 1969, *202*, 437–470.

Pribram, K. H. Some dimensions of remembering: Steps towards a neuropsychological model of memory. In J. Gaito (Ed.), *Macromolecules and Behavior*. New York: Appleton-Century-Crofts, 1966.

Pribram, K. H. The neurophysiology of remembering. *Scientific American*, 1969, *220*(1), 73–86.

Stroke, G. W. *An introduction to coherent optics and holography*. New York: Academic Press, 1966.

Westlake, P. R. *Towards a theory of brain functioning: A detailed investigation of the possibilities of neural holographic processes*. Unpublished doctoral dissertation, University of California, Los Angeles, 1968.

Willshaw, D. J. *Models of distributed associative memory*. Unpublished doctoral dissertation, Edinburgh University, 1971.

Willshaw, D. J. A simple model capable of inductive generalisation. *Proceedings of the Royal Society. Series B*, 1972, *182*, 233–247.

Willshaw, D. J. & Buneman, O. P. Parallel and serial methods of pattern matching. In D. Michie (Ed.), *Machine Intelligence 7*. Edinburgh University Press, 1972.

Willshaw, D. J., Buneman, O. P. & Longuet-Higgins, H. C. Non-holographic associative memory. *Nature*, 1969, *222*, 960–962.

Willshaw, D. J. & Longuet-Higgins, H. C. The holophone-recent developments. In D. Michie (Ed.). *Machine Intelligence 4*. Edinburgh University Press, 1969.