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 Unpublished doctoral dissertation, University of California, Los Angeles, 1968.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 Unpublished doctoral dissertation, Edinburgh University, 1971.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.

4 Storage and Processing of Information in Distributed Associative Memory Systems

Teuvo Kohonen
Erkki Oja
Helsinki University of Technology

Pekka Lehtiö
University of Helsinki

4.1. INTRODUCTION

4.1.1. A Classification of the Models of Memory Mechanisms

In traditional experimental psychology one of the principal goals of memory research has been to discover the quantitative laws that govern the performance of memory. A universal learning or forgetting curve may be seen as an ideal objective aimed at by researchers who are studying retention of words or letter strings in memory or learning of simple sensory-motor skills. The many stochastic learning models (e.g., Bush & Mosteller, 1955) published in the fifties and early sixties reflect only one facet of this general trend.

The most salient feature in contemporary memory research is the change of goal from the performance of memory to the mechanisms of encoding, retention, and retrieval of information. This gradual adoption of an idea that might be called *the memory mechanism paradigm* has also led to new research strategies. Work in this area has concentrated more and more on the modeling of memory functions; computer simulation models have become indispensable tools of theoretical work.

Although this common interest in memory mechanisms is recognizable in most current research, diversification of the goals has also led to a "balkanization" process (Newell, 1978) typical of different areas of cognitive science. For this reason one of the purposes of this introduction is to classify the various models of memory and to show how they have been motivated.

We divide memory models into two main categories: *physical system models* and *information-processing models*. In the category of *physical system models* we include all those models that try to answer the following question: How is it possible, using a collection of relatively simple elements connected to one another, to implement the basic functions of selective associative memory? Although it is possible to implement memory mechanisms by many different physical systems, we are here interested only in biological mechanisms for which there exists some experimental evidence and that have a plausible principle of operation. The fundamental question, dealt with in this chapter, concerns the mechanism that enables the nervous system to encode associations and subsequently to recall them selectively and independently. In order to explain the function of memory, the following three subproblems must be resolved: (1) What are the variable elements in the neural system capable of accumulating memory traces; (2) which neural events are identified with the reading and writing operations; and (3) what is the addressing mechanism used for memory in the nervous system?

The physical system approach may be contrasted with *information-processing models* that conceive man as an information-processing system executing internal programs for testing, comparing, analyzing, manipulating, and storing information. For an operable information processing paradigm, its propounders had to assume that the *mental processor* had some sort of associative memory capacity.

Within or at the fringe of the Artificial Intelligence community a great number of studies have been published on associative data structures. The aim of these studies has been to develop a representational format that permits the storage of the meaning of a word or a sentence, or more generally, the storage of organized knowledge. Quillian (1968) has formulated the central thesis of his early work on *semantic networks* as follows: "What constitutes a reasonable view of how semantic information is organized within a person's memory [p. 216]?" This is a typical competence question. A more technological approach was, of course, later adopted in the development of data-base software tools. Semantic network models and models of mental processing may be seen as complementary: For instance, in order to explain linguistic abilities, one has to postulate both processing functions and articulated data structures.

4.1.2. Representation of Information by Collective States

An inherent difficulty in the physical modeling of memory obviously arises from the fact that the human mind deals mostly with *concepts* that appear to be distinct and unique items. Attempts to construct abstract structures out of such hypothetical conceptual units are therefore understandable. Nonetheless, it has been

pointed out by Simon (1976): "Nothing in contemporary information-processing theories of memory requires that memories be specifically localized; and nothing in those theories is incompatible with a distributed or even holographic theory of physiological basis for memory [p. 80]." In this article an even stronger argument is presented. In fact, because a neural system is an ensemble of a great number of *collectively* interacting elements, it seems more natural to abandon altogether those physical models of memory in which particular concepts correspond to particular spatial locations (nodes) in the hardware. Instead a physically more plausible approach can be based on the assumption that representations of concepts and other pieces of information are stored as *collective states* of a neural network. It then becomes possible to demonstrate the formation of structured *interactions* between these distinct states. Structures of interactions can be made to correspond to *structures of knowledge;* however, they have no direct physical counterparts in the system. *They are realized only through the collective effects and reflected in recall processes.*

The first models of collective memory aimed at an analysis of the accuracy, capacity, and resolution achievable in the basic associative mappings, and they should not be misinterpreted on account of their simplicity. The principal objective in the early models was to show that there exist memory-dependent mappings between patterned sets of signals such that the memory traces can be superimposed on the same substrate in the form of some transformation functions without in any way losing information although the memory traces are superimposed on the same elements. The demonstration of the existence of such selective mappings showed that there may indeed exist distinct representations for pieces of information that nonetheless need not be represented and stored on the memory substrate spatially separated from each other. In the same way as many mathematical or physical entities can have functional expansions, for example, spectral decompositions, it thus became possible to demonstrate that distinct items can be represented as functional components spread all over the memory medium. This kind of collective representation might be called "holographic" or "hologic," although it need not resemble the usual optical holography that requires coherent wave fronts.

In the memory models advanced in this paper, the answers to questions concerning reading and writing of information, as well as the problem of addressing, follow from the idea that memory is regarded as an *adaptive filter*. When a neural network that implements this filter function is stimulated by signal patterns, each pattern produces adaptive changes in the neuronal interconnections, changes that are equivalent to the writing of information into the memory. Reading from memory can simply be the transformation of input signals in the network, because this transformation is dependent on the previous stimulus sequence. (Detailed description of these transformations as well as the adaptive changes is presented in Section 4.2.) No addressing problem exists because

memory traces are spatially distributed and superimposed throughout the network. This approach agrees well with current evidence about the functional structure of the neocortex (Creutzfeldt, 1976).

The very direct analogy between conceptual systems and neural structures was severely questioned at the beginning of the 1970s when several independent articles were published on collective effects as a basis of memory. We review in detail one of the most plausible models, the distributed associative network, in a form that seems to give rise to several fundamental information-processing paradigms in the neural realm.

4.2. ASSOCIATIVE MAPPINGS IN DISTRIBUTED MEMORY SYSTEMS

4.2.1. Associative Recall

A central operation in explaining the functions of distributed memory models is *associative recall*. As stated earlier, if the memorized data contains proper internal relations or links, structures of information can be shown to be reflected in the resulting recall processes, that is, without explicitly being represented in memory. Moreover it is possible to demonstrate processing of semantic data by associative recall as shown in Section 4.4.

In general terms the basic action of associative recall is definable as any process by which an input to the memory system, considered as a "key," is able to evoke in a highly selective fashion a specific *response,* associated with that key, at the system output. Associative recall implies a specific stimulus–response (S–R) type of *mapping* in the memory medium, which is able to associate a large number of large-scale activity patterns faithfully and also suppress errors. This mapping may be accompanied by many types of signal feedback and iterative operation.

The fundamental assumption in the approach based on associative mappings or transformations is that both the stimulus and the response are representable as complex, *patterned* sets of parallel signals. The mapping is not defined between individual signals but between these activity patterns as a whole, bringing together and interconnecting the various parts of the patterns. The results of this distributed or "hololog ic" mode of operation can exhibit many surprising features the existence of which is not obvious at first glance.

The models of distributed associative memory that were published first were primitive in the sense that only direct mappings between pictures or other simple patterns were demonstrated. However it was clear from the beginning that any complex signal representations, even with semantic contents and infrastructure, could have been chosen in place of the pictorial patterns used in simulations. In

other words, if the S–R mapping is definable for any sets of signals, then it is also possible to devise a mapping that transforms, for example, the representation of a *statement* into another one. Another fact to emphasize is that, quite intentionally, the S–R model was not loaded with auxiliary functions such as the activity-controlling projections in neural networks or preprocessing transformations of the patterns before they enter the proper memory system. By means of preprocessing operations it is possible to extract any type of features from the primary signals and to use them as a new basis of representation of information; this is already one step towards symbolism, although it too was deliberately neglected in the initial research.

In order to forestall misinterpretation of the network models of distributed memory presented in the following sections, it is necessary to point out that the whole brain is not assumed to be a single uniform network or "matrix" but a complex system consisting of many interacting parts. In the same way as a computer is made of chips of logic circuits, the brain may be composed of a great number of subunits, each one with the properties of a "memory matrix."

4.2.2. Two Structural Paradigms

As a starting point for concrete modeling of associative recall, consider the piece of network or hypothetical neural tissue depicted in Fig. 4.1 (Willshaw, Buneman, & Longuet-Higgins, 1969). The vertical units might represent the dendritic membranes of a set of neurons, and the horizontal lines could correspond to a set of axons or axon collaterals having synaptic connections on the dendrites.

The horizontal lines carry the elements of the *stimulus pattern* in the form of parallel, scalar-valued signals s_j. In a neural network the signal values would be represented by short-term averaged spike frequencies. At the output lines the vertical units send out the *responses* r_i, which similarly are represented by spike frequencies. In this idealized model, there is a *synaptic connection* m_{ij} between every vertical unit i and every horizontal line j. In practice some of the connections may be missing whereas others may be multiple. In the schematic representation the actual locations of connections on the units or the multiplicity of connections have not been shown explicitly.

In order that a given set of stimuli s_j evokes at the output another, predetermined set of response signals r_i, a selective S–R mapping must be encoded into the set of synaptic connections m_{ij}. One may call the array of m_{ij} values the "synaptic matrix." This encoding is adaptively and automatically formed when signals are mutually *conditioned* at the connections as explained below. Selective recall of a given response set from the synaptic matrix can thereby be rendered possible. The conditioned couplings may take many different functional forms. The simplest of them, the linear mapping, as explained in more detail in Section 4.2.3, is suitable as the first approximation. For this conditioning to happen, supervised (and *nonlinear*) learning must take place. For that purpose,

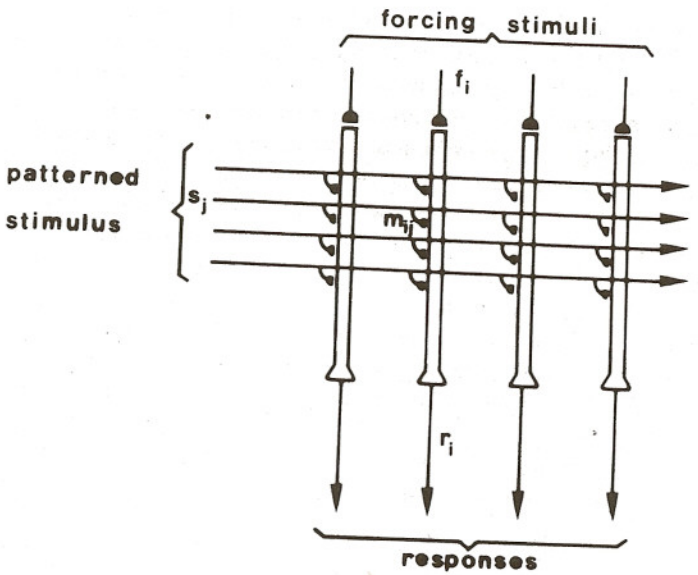

FIG. 4.1. Associative network with a set of connected neurons shown schematically. s_j = elements of the stimulus pattern; r_i = elements of the response pattern; f_i = elements of the forcing stimulus pattern, which are absent during recall; m_{ij} = synaptic connections.

the *forcing stimuli* have been introduced in Fig. 4.1. The forcing stimuli f_i appear as the primary input into the units. In this paradigm they are needed only when learning or *writing into memory* takes place; during learning the forcing stimuli have the same signal values as the desired output responses, which thereby become associated with the conditioning stimulus pattern. During recall there is no input at the forcing stimulus lines. A detailed analysis of the recall and the adaptive processes taking place in the above network, will be postponed to Section 4.2.3.

The system depicted in Fig. 4.1 is highly unsymmetric in the sense that the roles of the conditioning stimulus pattern and the forcing stimuli are strictly differentiated; in recall only the former type of input is used. This paradigm might already serve as a model for some neural structure like the cerebellum where the relation between the different inputs to the Purkinje cells is roughly of this kind. However, this model is unsatisfactory for modeling *cortical* regions of the brain. For instance, in the cerebral cortex there is a rich variety of recurrent activity mediated by axon collaterals and the longer subcortical projections. As a step towards a more realistic paradigm for cortexlike structures, one may replace the network of Fig. 4.1 by that given in Fig. 4.2.

The network of Fig. 4.2 shows two modular subsystems separated by vertical dotted lines. The forcing stimuli in this case comprise the primary patterned

input. Each subsystem is characterized by a dense net of interconnections; part or all of the output fibers are *fed back recurrently* into the cortical layer comprised of the parallel neuronal units, where the fibers branch and make redundant connections with the other units of the same subsystem. This simplified scheme neglects the existence of interneurons and other details of the actual topology of the connections; as shown later in Fig. 4.8, a substantial part of the short-range connections may be made subcortically. The network may also seem too homogeneous in that within a subsystem each unit is connected with all the other units. However, as shown in a previous work by one of the authors (Kohonen, 1972), all the connections do not in fact have to exist; if a portion of them is lacking, the system can still be statistically approximated by a complete set of connections.

In addition to short-range recurrent connections, there is a set of long-range connections from one subsystem to another. They introduce other inputs to every subsystem, which will become associated with the activity of the subsystem itself.

The essential feature in Fig. 4.2, referring to connections within a subsystem, is that in the synaptic matrix *every forcing stimulus is conditioned with all the other forcing stimuli of the same subsystem* through the recurrent connections. Within the subsystem there thus exists a complete symmetry with respect to the primary inputs. As a result it becomes possible to use any part of the forcing stimuli as a key. The activity pattern of the rest of the units is reconstructed in the

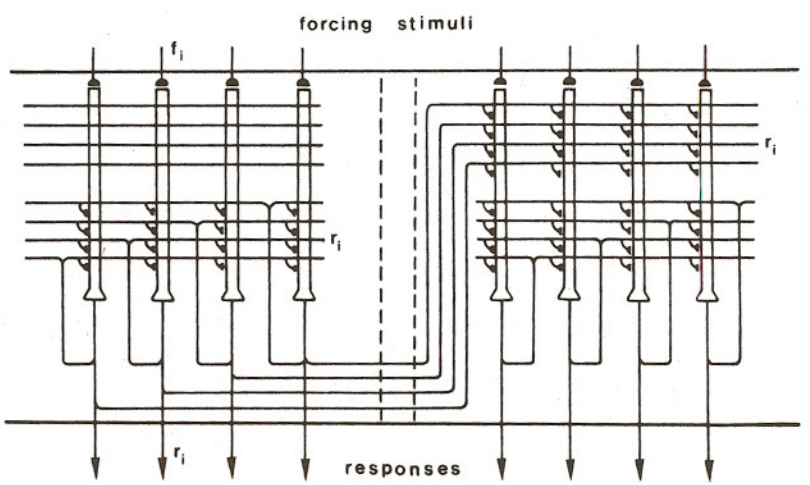

FIG. 4.2. The modular associative network with recurrent feedback. f_i = elements of the forcing or afferent stimulus pattern; r_i = elements of the response pattern that are fed back into the network at shorter and longer distances. The vertical dotted lines separate the two subsystems shown, and the horizontal solid lines represent the surfaces of the laminar network.

associative recall process through the interconnections. This is the *autoassociative encoding and recall principle* extensively advanced in this chapter.

The paradigms presented in Fig 4.1 and 4.2 represent the extreme and purest cases of S–R mapping, in the sense that Fig. 4.1 has no feedback whatsoever from the response back to the stimulus, but in Fig. 4.2 this feedback within one subsystem has the highest possible degree of completeness. The networks actually existing in the brain probably lie somewhere between these extreme paradigms, and thus their properties can be expected to be a mixture of the properties of these two networks.

In Fig. 4.3, the organization of Fig. 4.2 is shown as a three-dimensional structure. The slabs or subsystems separated by the dotted lines are the areas of the sheet with a high degree of interaction, but the interactions between the subsystems are weaker. The feedback connections mediating these interactions are no longer shown explicitly in Fig. 4.3. The organization depicted is the *laminar network model*, which is later used in the system-theoretical description of distributed memory.

The laminar network model is related (in more detail) to the anatomical and physiological properties of the mammalian brain in Section 4.3. There also, extra features, like the activity control exerted over selected areas, are merged into the model. A feature inherent in the model is that corresponding to the individual subsystems or regions in Fig. 4.2, the sets of input signals and output signals may also be organized into a number of subfields or parts having, for example, different modalities or semantic significance. The data contents of some or all of these subfields may also differ, and there may be neutral areas between them

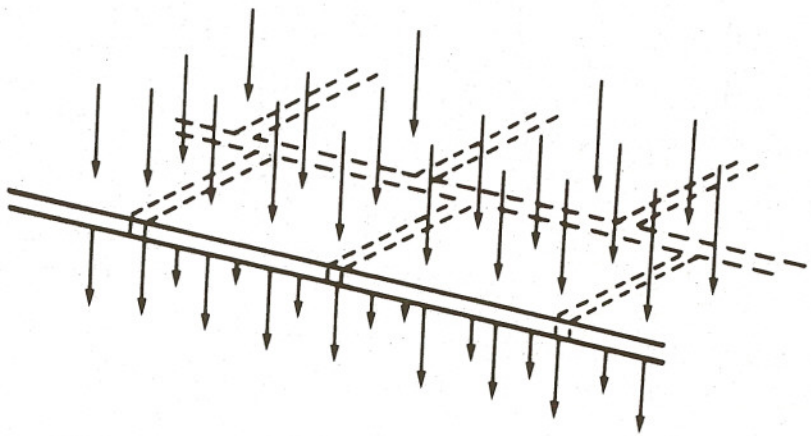

FIG. 4.3. Three-dimensional view of the modular system. An afferent stimulus pattern enters the top and a response leaves at the bottom. The dotted lines demarcate subfields or subsystems. The infrastructure of the lamina, not shown explicitly, corresponds to that given in Fig. 4.2.

with no signal activity (cf. Fig. 4.10). The important information-processing implications of the infrastructure in the patterns used in models of associative recall is postponed until Section 4.4. In order to describe information processing in such a model, as well as in the simpler associative network model presented earlier in Fig. 4.1, some quantitative analysis of signal transformations must first be presented.

4.2.3. Analysis of Adaptive Transformations

To describe as concretely as possible what kind of interactions between patterned data could be encoded into the networks in distributed form and how selective recall is achievable, we first return to the unsymmetric case of Fig. 4.1. The subsequent considerations of the network should be understood as a *system-theoretical approach* only; no assumptions will be made at this stage about the actual data represented by the patterns of activity.

As a first approximation to the transformations taking place in the network of Fig. 4.1, assume that each response signal, r_i, is a weighted sum of all the stimulus signals s_j and the ith forcing stimulus activity f_i,

$$r_i = \sum_j m_{ij} s_j + f_i, \tag{4-1}$$

where the weights m_{ij} stand for the synaptic conductivities. What was said previously about the necessity of a complete set of connections applies here, too: A portion of the connections m_{ij} may be lacking, and the network of Fig. 4.1 then serves as a statistical approximation.

It should further be stressed that the above linear mapping is only one representative in an infinite class of S–R transformations, some of which are derived from the linear models by adding nonlinearities like saturation or threshold triggering to the signal paths whereas others are nonlinear from the beginning. One possible source of misinterpretation of the earlier models of distributed associative memory was the assumed linearity of some approaches: Linearity was never intended to be an essential property but only a first approximation. Nonlinearities can also be introduced in the model as a separate operation, for example, at the output whereby the linear mapping can be assumed as the basic internal mode of operation of the system. In an alternative approach, the neural units are assumed inherently *binary* (Nakano, 1972; Willshaw et al., 1969); if the weighted sum of the binary-valued input signals exceeds a threshold, then the response is 1, otherwise 0. Despite the apparent nonlinearity of this functional form, the integral transformation properties are essentially the same as those explained here using the continuous linear approximation.

Although linearity does not seem to be a necessary property in signal transformation, the near-linearity of neural responses is often a quantitatively justified good approximation (Anderson & Silverstein, 1978).

Assume now that the strengths of the connections m_{ij} in Fig. 4.1 are changed during a learning phase by what is here termed *conjunctive forcing:* m_{ij} is changed only when both of the signals converging on it are active. There seems to be some physiological evidence, treated in Section 4.3, for this type of assumption involving two different signals. If changes (time derivatives) of the values of the m_{ij} are gradual, they are statistically described by a mathematical form in which the conjunction is replaced by the product of the signal values, resulting in the following correlation-type learning scheme:

$$\frac{d}{dt}\, m_{ij} = \lambda f_i s_j. \tag{4-2}$$

There λ is a scalar determined by the *plasticity* of the connections, which is here assumed constant for simplicity. It might also vary from one connection to another.

Let us now introduce the mathematical conceptualization of *patterns* into the above formalism. At a given instant, there are parallel stimulus signals, $s_1, \ldots, s_m$, on the input lines of Fig. 4.1. These make up a pattern $\mathbf{s} = (s_1, \ldots, s_m)$, simply defined as an ordered set of simultaneous parallel activities. In a similar fashion, at that same instant there is a forcing stimulus pattern, $\mathbf{f} = (f_1, \ldots, f_n)$, at the input lines and a response pattern, $\mathbf{r} = (r_1, \ldots, r_n)$, at the output lines. For economy of notation only (and without introducing any further mathematical or physiological assumptions) we shall now switch to *matrix algebra* (cf. Bellman, 1960) in the description of associative recall.

The synaptic conductivities m_{ij} can first be arranged into a rectangular array, forming a matrix $\mathbf{M}$ with n rows and m columns. The patterns **s, f,** and **r,** being sets of scalar numbers, can be identified with items in a high-dimensional vector space. The term vector here has an abstract meaning, totally different from the conventional two- or three-dimensional vectors used in geometry and field theory; here a vector is simply an ordered set of scalar numbers. It is hereafter assumed that **s** is a column vector of m dimensions, and both **f** and **r** are column vectors of n dimensions.

The pattern vectors **s, f,** and **r** and the matrix **M** can now be immediately substituted in Eq. (4-1) and (4-2), so (4-1) becomes

$$\mathbf{r} = \mathbf{M}\mathbf{s} + \mathbf{f} \tag{4-3}$$

and (4-2) becomes

$$d\mathbf{M}/dt = \lambda \mathbf{f}\mathbf{s}^{\mathrm{T}}, \tag{4-4}$$

with $\mathbf{s}^{\mathrm{T}}$ denoting the transpose of vector **s.**

Imagine now that there is a large number, say p, of different stimulus vectors. They are designated by $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(p)}$. Likewise, there are p different forcing stimulus vectors, $\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \ldots, \mathbf{f}^{(p)}$. The superscript is now used to separate between the different vectors; e.g., each $\mathbf{s}^{(i)}$ is composed of m elements that are

the individual parallel signal values making up the spatial activity pattern $\mathbf{s}^{(i)}$, whereas the corresponding m elements of $\mathbf{s}^{(j)}$ $(j \neq i)$, may be totally different from these.

Assume that all the synaptic strengths m_{ij} are initially 0, or $\mathbf{M}(0) = 0$; from the instant $t = 0$ onwards the first pair of stimulus and forcing stimulus vectors $(\mathbf{s}^{(1)}, \mathbf{f}^{(1)})$ appears at the inputs of the network of Fig. 4.1 and stays constant for a while. According to Eq. (4-4), matrix **M** then develops into

$$\mathbf{M}(t) = \lambda t\, \mathbf{f}^{(1)}\, \mathbf{s}^{(1)\mathrm{T}}. \tag{4-5}$$

From this time onwards, the second pair $(\mathbf{s}^{(2)}, \mathbf{f}^{(2)})$ appears and so on. If for convenience it is assumed that each stimulus pair is input to the network for a time period whose length is $1/\lambda$, then $\mathbf{M}(t)$ eventually develops into the matrix

$$\mathbf{M} = \sum_{i=1}^{p} \mathbf{f}^{(i)}\, \mathbf{s}^{(i)\mathrm{T}}. \tag{4-6}$$

Matrix **M** above takes the form of a *cross-correlation matrix.*

Based on this matrix, it is possible to recall associatively the forcing stimulus patterns using the primary stimuli as keys. It should be stressed that no constraints to the actual form and structure of the $\mathbf{f}^{(i)}$ patterns were given above; theoretically, any set of n scalar signals could appear there. In a concrete example, $\mathbf{f}^{(i)}$ could represent the *classification* of the corresponding stimulus pattern $\mathbf{s}^{(i)}$, in which case the classification of $\mathbf{s}^{(i)}$ would take place very simply by analyzing the response obtained when $\mathbf{s}^{(i)}$ has been used as the stimulus pattern. Of course any other pattern containing some type of information on the primary stimulus can be used as the associated forcing stimulus and then later be recalled associatively.

This becomes evident when the recall operation is defined according to Eq. (4-3) without, however, the forcing stimulus **f**, which was only necessary during the learning phase. Once the conjunctive learning phase described earlier is over, the specific values given in matrix form by Eq. (4-6) have been *imprinted* into the synaptic connectivities of the network. If now one of the earlier stimulus vectors, say $\mathbf{s}^{(j)}$, is used as the key stimulus in the network, the response becomes

$$\begin{aligned} \mathbf{r}^{(j)} = \mathbf{M}\,\mathbf{s}^{(j)} &= \sum_{i=1}^{p} \mathbf{f}^{(i)}\, \mathbf{s}^{(i)\mathrm{T}}\, \mathbf{s}^{(j)} \\ &= (\mathbf{s}^{(j)\mathrm{T}}\mathbf{s}^{(j)})\mathbf{f}^{(j)} + \sum_{i \neq j} (\mathbf{s}^{(i)\mathrm{T}}\, \mathbf{s}^{(j)})\mathbf{f}^{(i)}. \end{aligned} \tag{4-7}$$

In some cases different stimulus patterns have representations in terms of neural signals that can be assumed to be statistically independent. This independence is often expressable as a mathematical property named *orthogonality;* the key vectors are orthogonal if $\mathbf{s}^{(i)\mathrm{T}}\mathbf{s}^{(j)} = 0$ for $i \neq j$. Moreover, if signal values are

standardized, one can assume a metric property such that $\mathbf{s}^{(j)\mathrm{T}}\mathbf{s}^{(j)} = 1$. It then becomes evident that the response $\mathbf{r}^{(j)}$ is equal to $\mathbf{f}^{(j)}$: the response to $\mathbf{s}^{(j)}$ is an *exact recollection* of the forcing stimulus pattern $\mathbf{f}^{(j)}$ whose elements became conditioned with the elements of $\mathbf{s}^{(j)}$ in the course of learning. It must be emphasized that this is the *optimal* condition; the desired data are recalled completely and without any error. Because the maximum number of orthogonal m dimensional vectors $\mathbf{s}^{(j)}$ is m, this is also the maximum number of stimulus-response pairs that can be stored in the synaptic matrix of the network without violating the optimality condition.

If, however, the vectors used as keys above are not orthogonal, then the sum term in Eq. (4-7) will not be 0. It then represents *cross-talk* between the other stored patterns, and the less orthogonal the key patterns are in general, the higher is the cross-talk as compared to the correct recollection.

The occurrence of cross-talk that seems to limit the memory capacity has led to an interesting theoretical question: One may ask whether it is possible to devise a network of the above kind that would implement associative recall with *ideal selectively,* that is, in which, with a hypothetical synaptic matrix $\mathbf{M}$, the desired stimulus-response relation would be implementable for *arbitrary* pairs of patterns $(\mathbf{s}^{(j)}, \mathbf{f}^{(j)})$ such that

$$\mathbf{f}^{(j)} = \mathbf{M}\,\mathbf{s}^{(j)} \quad \text{for all } j. \tag{4-8}$$

Although information processing implementable by neural functions might quite well employ the powerful property of orthogonality, it is intriguing to find that this problem has solutions that are independent of the orthogonality assumption. This is one of the basic problems studied in linear algebra, and it has a simple answer that can be expressed in the form of a theorem:

Theorem. If all the $\mathbf{s}^{(j)}$ are linearly independent (no one can be expressed as a linear combination of the others), then a solution of Eq. (4-8) exists and is given by

$$\mathbf{M} = \mathbf{F}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}} \tag{4-9}$$

where $\mathbf{F} = (\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)})$ and $\mathbf{S} = (\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(p)})$ are the matrices with the $\mathbf{f}^{(j)}$ and $\mathbf{s}^{(j)}$ as their columns, respectively, and the superscript T denotes the transpose of a matrix. If the vectors $\mathbf{s}^{(j)}$ are not linearly independent, then there exists a unique approximative solution in the sense of least squares

$$\hat{\mathbf{M}} = \mathbf{F}\mathbf{S}^{+} \tag{4-10}$$

where $\mathbf{S}^{+}$ is the *pseudoinverse* of $\mathbf{S}$ (Albert, 1972). If vectors $\mathbf{s}^{(j)}$ are linearly independent, then in fact $\mathbf{S}^{+} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}$; see Eq. (4-9).

Incidentally, $\hat{\mathbf{M}}$ has the form of the *best linear unbiased estimator* (BLUE), which is a kind of Gauss-Markov estimator (Lewis & Odell, 1971). Theoretically, even if there were an infinite number of pairs $(\mathbf{s}^{(j)}, \mathbf{f}^{(j)})$, but they were *clustered,* there would nonetheless exist an approximate solution $\hat{\mathbf{M}}$ which defines an "infinite" associative memory in the sense of Eq. (4-8). Matrices of Eq.

(4-9) and (4-10) represent the *optimal linear associative mapping* that has been extensively studied previously by one of the authors (Kohonen, 1977). A demonstration of the use of the optimal linear associative mapping in classification of pictorial patterns is shown in Fig. 4.5 of Section 4.2.7.

4.2.4. Autoassociative Encoding

Very interesting associative recollections are obtained if the optimal linear associative mapping is considered in the case $\mathbf{f}^{(i)} = \mathbf{s}^{(i)}$. Of course, the trivial solution of Eq. (4-8) is $\mathbf{M} = \mathbf{I}$ (identity matrix), but this has no sense; one should set up a more general solution which according to Eq. (4-10) reads

$$\mathbf{M} = \mathbf{F}\mathbf{F}^{+} \tag{4-11}$$

Incidentally, this is a so-called *orthogonal projection operator* or *projector* with some interesting pattern-processing properties.

The p different vectors, $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$, all of them n-dimensional, span a *linear subspace* $\mathcal{L}$ in the n-dimensional vector space, that is, the vectors constitute a basis of $\mathcal{L}$. In still other words, $\mathcal{L}$ is the set of vectors that results from all possible linear combinations of the $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$. It is a well-known result from the theory of Hilbert spaces that an arbitrary pattern vector $\mathbf{f}$ with dimensionality n can always be uniquely decomposed into a sum of two component vectors

$$\mathbf{f} = \hat{\mathbf{f}} + \tilde{\mathbf{f}} \tag{4-12}$$

such that $\hat{\mathbf{f}}$, obtained by

$$\hat{\mathbf{f}} = \mathbf{F}\mathbf{F}^{+}\mathbf{f} = \mathbf{M}\mathbf{f} \tag{4-13}$$

is the *linear regression* of the $\mathbf{f}^{(j)}$ on $\mathbf{f}$ or the best linear combination in terms of least squares, and $\tilde{\mathbf{f}}$ is the residual. In fact, $\hat{\mathbf{f}}$ is contained in the subspace $\mathcal{L}$ whereas $\tilde{\mathbf{f}}$ is orthogonal to $\mathcal{L}$; hence the two vectors are mutually orthogonal, too. We can call $\hat{\mathbf{f}}$ the *optimal autoassociative recollection* relative to the stored information, or the set $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$, and the search argument or key pattern $\mathbf{f}$. Similarly, the matrix $\mathbf{M}$ of Eq. (4-11) is the *optimal linear autoassociative mapping,* which in spite of its simple form will be seen to be capable of processing patterns in rather unexpected ways.

The orthonormality of the stored patterns $\mathbf{f}^{(j)}$ would allow the matrix of Eq. (4-11) to be reduced to another form. For pattern vectors $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$ such that $\mathbf{f}^{(i)\mathrm{T}}\mathbf{f}^{(j)} = 0$ for $i \neq j$ and $\mathbf{f}^{(j)\mathrm{T}}\mathbf{f}^{(j)} = 1$, the corresponding projection operator reads

$$\mathbf{M} = \mathbf{F}\mathbf{F}^{T} = \sum_{i=1}^{p} \mathbf{f}^{(i)}\,\mathbf{f}^{(i)\mathrm{T}}. \tag{4-14}$$

The matrix above has the form of an *autocorrelation matrix.* Assume now that a new (independent) pattern vector $\mathbf{f}$ is given as a key input excitation. Its component vectors $\hat{\mathbf{f}}$ and $\tilde{\mathbf{f}}$, can be presented in the form

$$\hat{\mathbf{f}} = \mathbf{M}\mathbf{f} = \sum_{i=1}^{p} (\mathbf{f}^{(i)T}\mathbf{f})\mathbf{f}^{(i)}, \tag{4-15}$$

$$\tilde{\mathbf{f}} = \mathbf{f} - \hat{\mathbf{f}}. \tag{4-16}$$

Eq. (4-15) shows clearly how $\hat{\mathbf{f}}$, or the optimal autoassociative recollection, is now obtained as a linear combination of the stored vectors, where the coefficients of this expansion are simply the inner products of the stored vectors with the new vector **f**.

Even in the case of nonorthogonal vectors there exists a mathematical relationship between the optimal autoassociative mapping and the autocorrelation matrix; when matrix $\mathbf{FF}^{+}$ is presented in the form of a von Neumann expansion (cf. Rao & Mitra, 1971)

$$\mathbf{FF}^{+} = \alpha \sum_{k=0}^{\infty} \mathbf{FF}^{T}(\mathbf{I} - \alpha\mathbf{FF}^{T})^{k} \tag{4-17}$$

where **I** is the identity matrix, and the matrix series on the right is written out, then the matrix $\alpha\mathbf{FF}^{T}$ appears as the 0-th degree term. There α is a scalar that lies between predetermined bounds. In this sense, the autocorrelation matrix may be regarded as a 0-th degree approximation of the optimal mapping.

Just as the cross-correlation matrix was imprinted into the synaptic connections of the associative network of Fig. 4.1 by conjunctive forcing, so the autocorrelation matrix, Eq. (4-14), can be shown to be the outcome of a similar learning process in the feedback network of Fig. 4.2. If the modular organization there is approximated by a homogeneous set of mutual feedback connections between the units, then the main difference between mathematical considerations of this network and the one of Fig. 4.1 is that, due to the feedback, the response signals r_i now appear in place of primary signals s_i. This difference becomes clear when the two figures are compared. One horizontal line in Fig. 4.2 carries one response signal r_i, but a corresponding horizontal line in Fig. 4.1 carries an external stimulus signal s_i.

Because of the feedback, however, the signal transmission properties and their mathematical treatment are not as straightforward in this case as in the paradigm considered in the previous Section. This problem has been considered in detail by Kohonen, Lehtiö, Rovamo, Hyvärinen, Bry, and Vainio (1977). First of all, in order to explain the adaptive formation of the optimal linear autoassociative mapping in the laminar network model, one has to take into account *short-range lateral inhibition* between the units or columns. This has the effect of forming weighted sums of neighboring activities locally; in terms of the forcing stimuli f_i appearing at the top of the laminar network, each signal f_i should be replaced by an "effective excitation"

$$f_i' = \sum_{j} \beta_{ij} f_j \tag{4-18}$$

where the weights β_{ii}, corresponding to direct connectivities, are positive, but the lateral connectivities β_{ij} with $i \neq j$ are predominantly negative or inhibitory; index j runs over a surround of unit i. The numerical values of β_{ij} thus reflect the excitatory and inhibitory penumbrae around a point of afferent excitation. An interesting consequence of lateral inhibition is that, for typical two-dimensional pictorial patterns, the transformation produced by Eq. (4-18) has the effect of orthogonalizing the patterns fairly effectively.

When the orthogonalized input patterns enter the laminar network one at the time and conjunctive learning takes place, the network connections are adaptively changed. Later, when an input pattern is applied, the signals are modified by the network, due to the adaptive effects caused by the earlier signals. This is equivalent to the *reading of stored information from memory*. For details, see Kohonen et al. (1977a).

The autoassociative mapping has the ability to reconstruct any of the stored patterns when only a part of the pattern or a distorted version (e.g., contaminated with noise) is used as the key input. Some demonstrations are shown in Section 4.2.7. There are more far-reaching properties, too, which can be best explained by the subspace formalism of Eq. (4-12). In fact even when the key pattern **f** bears no similarity whatsoever to any of the previously stored patterns, $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$, the output or optimal autoassociative recollection $\hat{\mathbf{f}}$ must still reveal some characteristic features common to all the stored patterns, because $\hat{\mathbf{f}}$ is always a vector in the subspace $\mathcal{L}$ spanned by $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$. This implies synthesizing and generalizing abilities, further illustrated by a demonstration in Section 4.2.7.

4.2.5. Extraction of Novelty

In the subspace formalism just presented, it was stated that the projection vector $\hat{\mathbf{f}}$ represented the recollection from memory, and the orthogonal component $\tilde{\mathbf{f}}$ then assumed the role of a *residual*. In regression analysis, the residual would be assumed noiselike and be inversely related to the goodness-of-fit; in the present considerations, however, vector $\tilde{\mathbf{f}}$ is better understood as the result of a particular information-processing operation whose purpose is to *filter out* from the pattern vector **f** the component that is explained by the stored data. It is then possible to think of $\tilde{\mathbf{f}}$ as the amount that is "maximally new" in **f**. It may be justified to call this component the *novelty with respect to the stored vectors*, and the name *Novelty Filter* has been used for a system which extracts $\tilde{\mathbf{f}}$ from input data **f** and displays it at the output alone without the $\hat{\mathbf{f}}$ component (Kohonen, 1977; Kohonen & Oja, 1976). The Novelty Filter system has the ability to enhance any nonfamiliar part or features appearing in an activity pattern passing through.

Because $\hat{\mathbf{f}}$ is obtained from **f** by a linear operation, the application of the projection matrix $\mathbf{FF}^{+}$, where **F** is the matrix whose columns are the stored vectors $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$, then it follows immediately that

$$\tilde{\mathbf{f}} = \mathbf{f} - \mathbf{FF}^{+}\mathbf{f} = (\mathbf{I} - \mathbf{FF}^{+})\mathbf{f}. \tag{4-19}$$

where $\mathbf{I}$ is the identity matrix. This shows that $\tilde{\mathbf{f}}$ is the outcome of a linear operation, too, with the matrix

$$\mathbf{P} = \mathbf{I} - \mathbf{F}\mathbf{F}^{+} \tag{4-20}$$

giving the linear operator. This matrix is the transfer operator of the Novelty Filter. It is a projection matrix, too, because it projects every vector on a subspace $\mathcal{L}^{\perp}$ which is the orthogonal complement of the subspace $\mathcal{L}$ spanned by the stored vectors $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(p)}$.

A network implementation for the Novelty Filter paradigm can be explained in terms of the laminar model with internal feedback, Fig. 4.2. If the separate afferent input signals of the network are again denoted by scalars f_i, the response signals by r_i, and the connectivities of the network by m_{ij}, then the output signals or responses are actually functions of both the afferent input and the other responses of the network, carried by the feedback lines converging on the units. By exact analogy to the mathematical treatment performed earlier in Section 4.2.3., with the same comments applying for the linearity approximation and the completeness of feedback connections, we then have

$$r_i = f_i + \sum_j m_{ij} r_j. \tag{4-21}$$

In vector-matrix form, this reads

$$\mathbf{r} = \mathbf{f} + \mathbf{M}\mathbf{r}, \tag{4-22}$$

which further yields

$$\mathbf{r} = (\mathbf{I} - \mathbf{M})^{-1}\mathbf{f}. \tag{4-23}$$

Thus the *overall* transfer operator for the input patterns is in fact matrix $\Phi = (\mathbf{I} - \mathbf{M})^{-1}$. The crucial difference with respect to the earlier autoassociative network lies in the assumed law for synaptic modification, which in place of Eq. (4-2) now reads

$$\frac{d}{dt} m_{ij} = -\alpha r_i r_j. \tag{4-24}$$

In other words, *negative feedback* (mainly due to inhibitory connections) has been introduced, and it will build up adaptively, tending to compensate for the input excitation. Actually the physical implementation of Eq. (4-24) has some extra details not discussed here (Kohonen & Oja, 1976). The laminar model, described by the transfer operator Φ, now becomes equivalent to a very special and selective "habituating filter"; it will display at its output only that component of the input pattern vector which is orthogonal to the subspace already spanned by all earlier inputs, that is, it displays exactly the previously mentioned residual. If each input is applied for a suitably long time, the resulting transfer matrix is in fact $\Phi = \mathbf{P}$, the Novelty Filter projector given by Eq. (4-20). The

mathematical form that the learning process takes is describable by a Bernoulli matrix differential equation, whose solutions are intimately related to an algorithm well known in linear algebra, viz. the Gram-Schmidt orthogonalization procedure (Albert, 1972; Oja, 1978). Therefore the term *orthogonalizing filter* has also been used for this system.

Novelty Filtering is a useful fundamental operation in any natural information storing and processing system, because those parts or features of a pattern that are directly expressible in terms of the stored data are rejected, but the filter system is transparent to the most interesting components of new data.

4.2.6. Recollection of Temporal Sequences

The purpose of this section is to demonstrate that *temporal associations* are easily obtainable from an associative memory provided with minor extra features, namely, *delayed feedback*.

Imagine that a sequence of input patterns $\{\mathbf{A}(t)\}$ that all share the same *background,* or *context,* **B** shall be stored. Context **B** may now be regarded as a specific part of the input field; however a characteristic of signals applied in this part is that *they are always held constant during a particular sequence* $\{\mathbf{A}(t)\}$. On the other hand, different context signals can be used for different sequences. Due to the introduction of context the same pattern $\mathbf{A}(t)$ may thus be associated with several different output patterns using different contexts. The role of context has become very central in the temporal mode of operation; by its virtue it becomes possible to identify the different sequences directly and to select one of them for recall.

Consider the sequential machine depicted in Fig. 4.4 (Kohonen, 1977, 1980). The central block is some kind of autoassociative distributed memory network, for example, a laminar memory model discussed earlier. The system receives two types of input: the external input, consisting of **A** and **B,** and feedback input **D.** The feedback is obtained from the output of the system through a delay that, for simplicity, is assumed to have unit length. It may be assumed that if **A** represents a forcing input, then during the input process (writing into memory) **C** is a response that is a replica of **A.** During recall, however, there is no forcing input and **C** is recalled by (**B, D**) used as the key. Assume now that a sequence of inputs $\{[\mathbf{A}(1), \mathbf{B}], [\mathbf{A}(2), \mathbf{B}], \ldots, [\mathbf{A}(N), \mathbf{B}]\}$ is received and stored autoassociatively. When the first input pattern $[\mathbf{A}(1), \mathbf{B}]$ arrives, the feedback input **D,** due to delay, does not receive any signals yet from **C;** it consists of an empty subpattern ϕ. When the second pattern $[\mathbf{A}(2), \mathbf{B}]$ arrives, the previous output $\mathbf{A}(1)$ has already been transmitted through the delay, and it is assumed to appear synchronously with $\mathbf{A}(1)$. According to the above considerations, the *effective* input sequence of the autoassociative memory is then

$$S = \{[\mathbf{A}(1), \mathbf{B}, \phi], [\mathbf{A}(2), \mathbf{B}, \mathbf{A}(1)], \ldots, [\mathbf{A}(N), \mathbf{B}, \mathbf{A}(N\text{-}1)]\}. \tag{4-25}$$

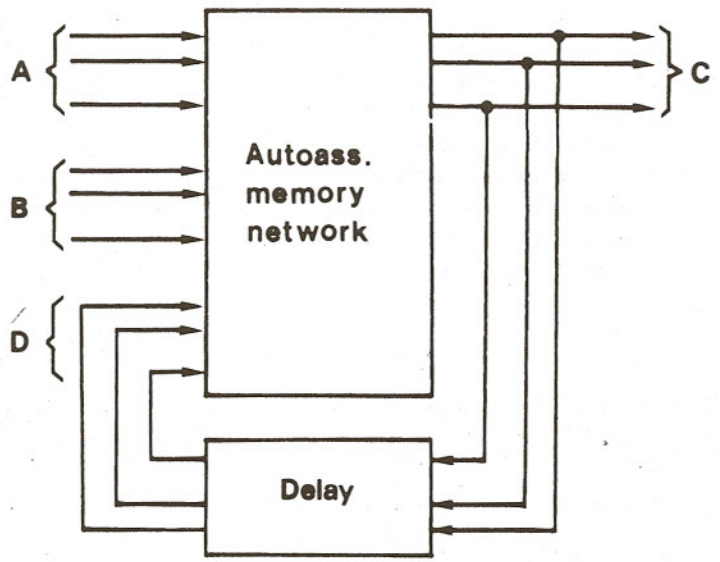

FIG. 4.4. A system for the associative recall of sequences. **A** = forcing input vector; **B** = constant background or context pattern; **C** = response pattern, the recollection from autoassociative memory; **D** = feedback pattern, equal to the response at a previous instant, with the time difference given by the delay.

In order to recall the $\{\mathbf{A}(t)\}$ sequence, the reading is started at time $t = 0$ by applying a key input $[\mathbf{A}(1), \mathbf{B}, \phi]$. Hereafter, no other members of the sequence are needed because they are automatically produced by the memory. At $t = 1$, the input pattern is now $[\phi, \mathbf{B}, \mathbf{A}(1)]$ where $\mathbf{A}(1)$ appears because of the feedback. The memory recalls the missing part $\mathbf{A}(2)$ of this activity pattern associatively and produces it at the output. In this manner a continued autonomous process will retrieve the whole sequence $\mathbf{A}(2), \mathbf{A}(3), \ldots, \mathbf{A}(N)$ associated with the background **B**.

A problem arises if the same state occurs in several places in the sequence and it has a different successor state each time. It should then be realized that the system of Fig. 4.4 is the simplest model containing feedback loops. If there are several feedback paths with different delays, then the model will be able to recall more complicated sequences; in fact, a machine with k different feedback paths is needed to recall correctly a sequence that contains several identical subsequences of length $(k\text{-}1)$.

4.2.7. Some Demonstrations of the Pattern Processing Properties of Optimal Associative Mappings

The simple demonstrations reported in this section are only intended to illustrate pattern-transforming effects of the basic optimal associative mappings; processing of structured information needs a more developed organization as delineated, for example, in Section 4.4. Nor is it claimed that the well-known problem of invariances in perception would be completely solved by the interpolation that takes place in linear transformations. However, even though this experiment does not simulate the operations of the complete visual system, the pictorial material (human faces) is justifiable as test data because it has an inherently natural statistical structure and allows direct inspection of the recollections. The actual neural patterns would look quite different because they are transmitted through many preprocessing stations.

Pattern Classification. As a first demonstration, consider a classifier based on the optimal linear associative mapping for pattern pairs $(\mathbf{s}^{(j)}, \mathbf{f}^{(j)})$; see Eq. (4-9). Here the $\mathbf{s}^{(j)}$ are *prototype* vectors that belong to various *classes;* each $\mathbf{f}^{(j)}$ is a unit vector that has as many components as there are classes of patterns. To every class there corresponds a unit vector with the value 1 in a particular component and 0 elsewhere. For every class, a small number of prototypes is collected; if $\mathbf{s}^{(j)}$ is one of the prototypes, then

$$\mathbf{f}^{(j)} = \mathbf{M}\,\mathbf{s}^{(j)} \quad \text{for all } j \tag{4-26}$$

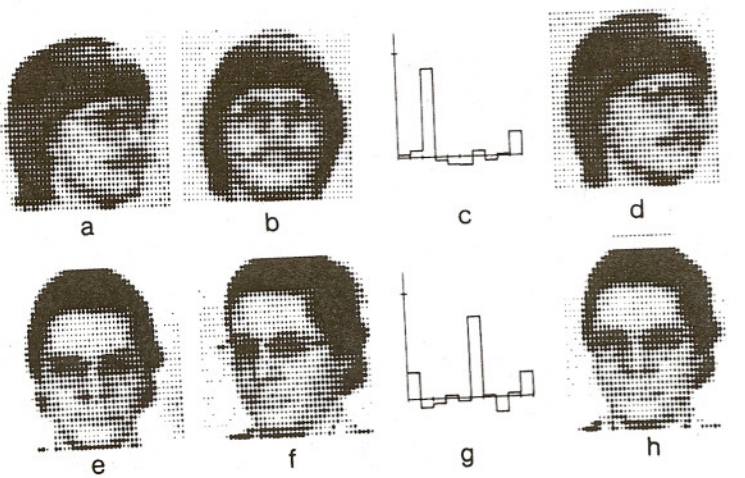

FIG. 4.5. Demonstration of classification by optimal associative mapping. Each of the 10 pattern classes employed consisted of pictures of one person photographed from five different angles, ranging from +45° to −45°. Image vectors with components consisting of discrete picture elements were used as pattern vectors; eight intensity levels were defined for each picture element. A distinct unit vector was associated with each person. Parts (a) and (b) show two prototypes from one pattern class (no. 3), and Parts (e) and (f) show two prototypes from another pattern class (no. 6). Part (d) shows a test image of the person in (a) and (b), taken from an angle not used among the prototypes. In the histogram of the recollection, (c), the position of the largest component correctly reveals the number of the class. Parts (g) and (h) repeat the same with another class.

holds exactly. If, however, a vector $\mathbf{s}$ to be classified has only varying degrees of similarity with the prototypes of different classes, the matrix–vector product $\mathbf{Ms}$ (associative mapping) produces an output vector $\mathbf{f}$ that has the same dimensionality as the unit vectors $\mathbf{f}^{(j)}$ but in which each component only describes a "weight" by which $\mathbf{s}$ belongs to the various classes. The largest component, or weight, is assumed to indicate the classification.

In the demonstration of Fig. 4.5, 10 persons were viewed from different angles, and these images were used as the prototypes of 10 classes, each class corresponding to one person (Kohonen, Lehtiö, Oja, Kortekangas, & Mäkisara, 1977). These prototypes were associated with unit vectors corresponding to different persons. An associative mapping of the images onto unit vectors is seen to be able to interpolate between the representations so as to yield a correct recognition of a person from a viewing angle not occuring in the contents of the memory.

Autoassociative Recall. In order to demonstrate the autoassociative mapping given in Eq. (4–11), the patterns $\mathbf{f}^{(j)}$ that were stored in memory were facial images of different persons; there were 100 pattern vectors stored ($p = 100$). Four of the images are shown in Fig. 4.6 (a-d). An incomplete or noisy version of

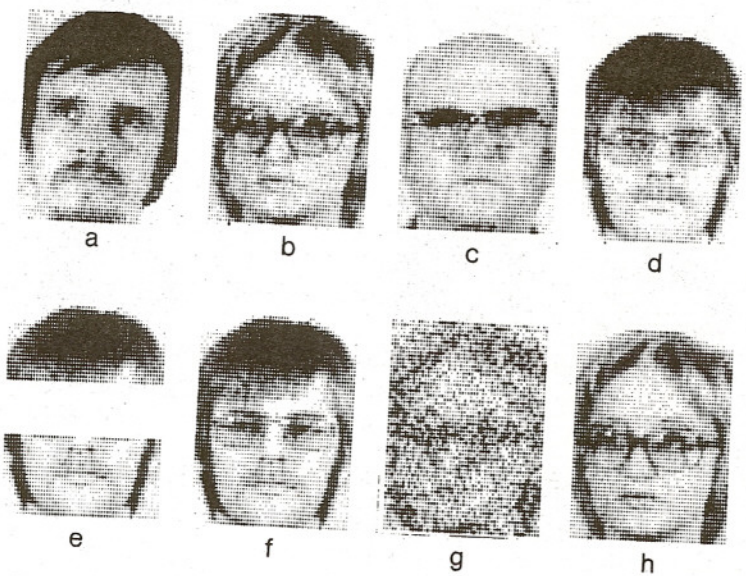

FIG. 4.6. Demonstration of autoassociative recall. Parts (a) through (d) show 4 of the 100 prototype images used to construct the autoassociative projector. When an incomplete or noisy version of a prototype, (e) and (g), respectively, served as the key pattern, the recollection resulting in the optimal autoassociative mapping is then shown to reconstruct the original appearance in (f) and (h), respectively.

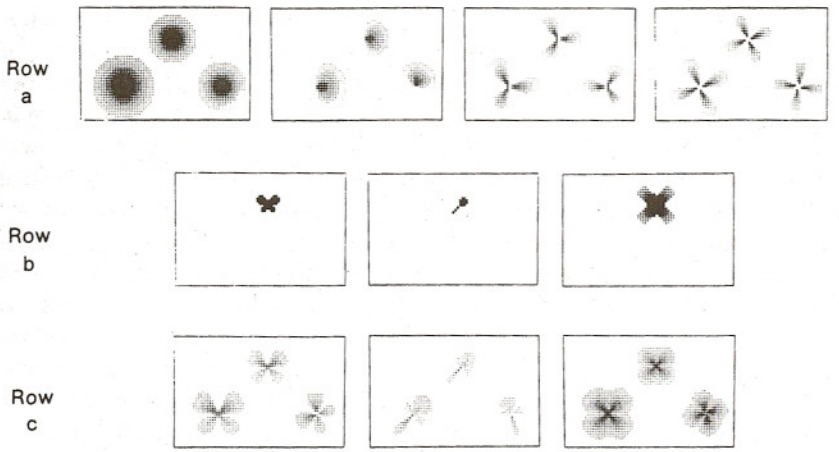

FIG. 4.7. Demonstration of generalization and synthesis by an autoassociative mapping. Row a shows 4 of the 21 different prototypes, each one composed of 3 subpatterns with invariant relations of size, location, and orientation. Row b displays new subpatterns never occurring among the prototypes. Each subpattern is located in the same place as the uppermost subpattern in the prototype images, but the rest of the picture field is empty. In Row c each pattern is the recollection from the autoassociative mapping, computed from the 21 prototypes, when the pattern above it was used as the key. The autoassociative mapping is shown to synthesize new subpatterns exhibiting the same invariance of size, location, and orientation as the prototypes.

one of the stored images, Fig. 4.6(e) or 4.6(g), respectively, is now taken as the key $\mathbf{f}$. The optimal recollection or projection $\hat{\mathbf{f}}$ is shown to reconstruct the original appearance, at least approximately, in Fig. 4.6(f) and 4.6(h), respectively.

Another demonstration of autoassociative mapping, emphasizing its synthesizing and generalizing properties, is shown in Fig. 4.7. The prototype patterns were defined as two-dimensional functions on the image field. The field consisted of 3128 points, but in this experiment only 21 different stored patterns were used. Every stored pattern consisted of three subfields each containing a subpattern, whose relations of location, size, and orientation were *invariant* in all prototypes. The *microstructure* in each subpattern, expressed in polar coordinates, consisted of bell-shaped functions in the radial direction and sinusoidal functions with varying frequency and phase in the angular direction. The subpatterns of each prototype were so constructed that the prototype patterns, considered as 3128 component vectors (read from the picture field column by column, with vector elements being the gray scale values of the picture elements), were all linearly independent, in fact orthogonal. Orthogonality is, however, not necessary in principle, but it simplifies computing algorithms because the autocorrelation matrix (cf. Eq. (4–14)) can now be used instead of the autoassociative mapping.

Four of the 21 stored prototype patterns are shown in Row *a* of Fig. 4.7. Various test patterns were used as key patterns, each consisting of some form of subpattern in the same location as the uppermost subpattern in all the prototypes but with the rest of the picture field empty. Three such test patterns are seen in Row *b* of Fig. 4.7. The results of the demonstration, the recollections in Row *c* of Fig. 4.7, show that the invariances in the stored patterns, that is, their internal structure, have become an inherent property of the space of linear expansions and are thus implicitly contained in the elements of the memory mapping. The recollections exhibit in every case the same relations of size, location, and orientation as the prototypes but with the three subpatterns in the recollection similar in form to the one in the key. Even with such a small set of basis functions the autoassociative mapping has achieved the ability to synthesize the two new subpatterns, resembling in form the one used in the key, and also to generalize the structure of the prototypes to hold between these subpatterns of the recollection. It might be expected that with a larger and more representative set of prototypes the recollections would comply even better to the key patterns.

4.3. THE NEURAL IMPLEMENTATION OF ASSOCIATIVE MEMORY

4.3.1. The Laminar Model

Distributed associative memory seems to be implemented in the brain as laminar networks with internal and sublaminar connections. In a fairly homogeneous lamina, it is easy to see how the operation of the network might be mathematically described by a matrix, which may be sparsely occupied.

How does this simplified system-theoretical view accord with the physiological and anatomical reality of the complicated neural machinery comprising the mammalian brain, and what might be the neural embodiment of such a distributed adaptive memory system? These are questions that should be answered before trying to explain mental information processing by the distributed associative paradigm. First of all, the histological analysis of the brain lends support to the idea that neurons are organized into horizontal sheets. This laminar organization is found in neocortex, allocortex, cerebellum, and in many areas of midbrain. (For an extensive collection of empirical results on different laminated structures, see Creutzfeldt, 1976.) The recent findings on these histological structures seem to point to a type of functional organization, whose overall behavior might be approximately described by the mathematical apparatus reviewed earlier. The structure of neocortex is used in the following description as a prototype of all cortical laminated structures.

Although it is morphologically possible to distinguish several layers in the cortical lamina, physiological studies on the mammalian cortex have revealed

that the responses are similar in all cells that are confined within vertical columnlike aggregates of cells extending from pia to white matter (Hubel & Wiesel, 1974; Mountcastle, 1957). It is generally accepted that such columns (or slabs) are organized around specific afferent axonal inputs, which they seem to analyze. Teleologically we may think that the columnar organization arises from the necessity to represent different stimulus qualities upon a two-dimensional surface, at the same time preserving the topological organization. Consider the schematic view of cerebrocortical modular organization given in Fig. 4.8(a). This view has been grossly simplified by presenting the net as a "skeleton cortex," omitting most of the cell types and taking into account only the prin-

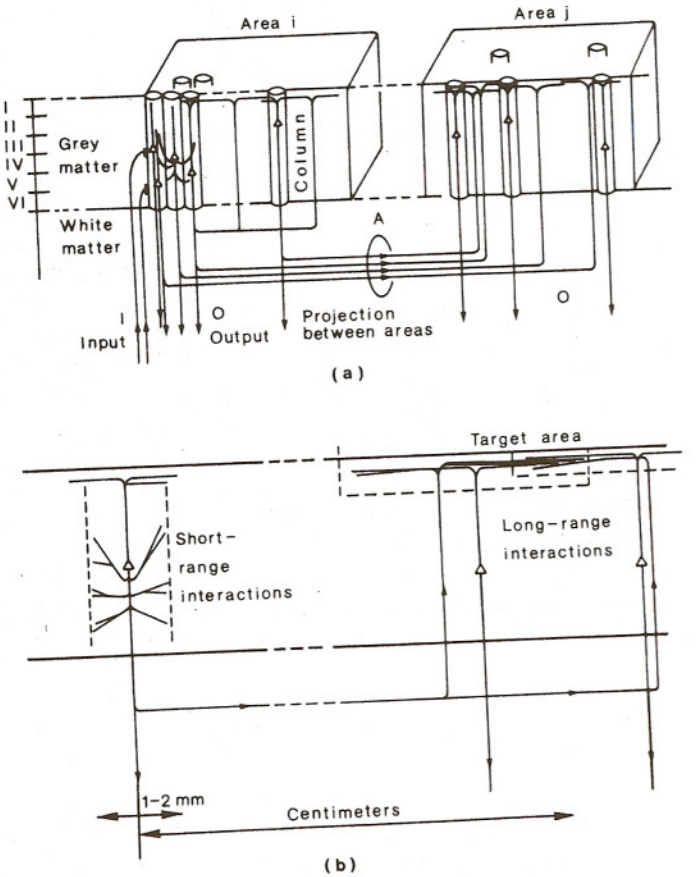

FIG. 4.8. (a) Cortical modular organization. I = afferent inputs, with actual termination not shown explicitly; O = outputs; A = association or commissural fibers mediating a projection between cortical areas. b) Short- and long-range interactions.

cipal neurons, which are pyramidal cells (Shepherd, 1974). The two block-like structures depicted in Fig. 4.8a are cortical modules or areas defined by their cortico-cortical connections. These connections are mediated by the subcortical axons of the pyramidal cells, which, after passing the white matter, ascend vertically through the cortical lamina up to its uppermost two layers, thereby ramifying with a great number of branches and extending over an area with a diameter of 2–3 mm (Szentágothai, 1978). Pyramidal cells have a similarly branching tree of apical dendrites in these layers, which allows a high degree of horizontal interconnectivity between cells. As every pyramidal cell has further axon collaterals that branch within the cortex, one may distinguish between the following three types of horizontal connectivity: (1) intracortical excitatory connections made by axon collaterals at a distance up to 100 μ; (2) intracortical inhibitory connections, possibly mediated through interneurons, which extend to a distance of 500 μ, and depend strongly on distance; and (3) cortico-cortical excitatory connections that may extend over any distance within the cortex. If these connections are made within the same area, then connectivity is independent of distance, and its obvious purpose is to provide the nonspecific interactions between cells. If these cell connections project from one area to another they are scattered over the target area. The commissural fibers between the cortical hemispheres also belong to the class of long-range connections but with specific connectivity. The two latter types of interaction are further shown schematically in Fig. 4.8(b).

It is assumed that the connections of type (3) carry out the memory-dependent interaction of columns. The integrated memory effects are therefore mediated by apical dendrites, which are known to have weaker but numerous excitatory synapses (Shepherd, 1974). The synaptic conductivities m_{ij} (Section 4.2) are used to describe the behavior of these contacts.

The spatial spread of the apical dendrites and the axons terminating at layer I increases the number of synapses and thus ensures a high degree of connectivity. It should be noticed that even if each column is integrating the activity of an area far beyond its dimensions, the response is generated locally by the column itself. There is therefore no lack of resolution in the output activity of the network.

4.3.2. Processing of Information by Interconnected Cortical Areas

It is proposed here that information is processed in interconnected cortical areas. This type of organization emphasizes the fact that each column may have inputs from different processing levels or from different auxiliary areas. Although it is possible to have a sequence of processing levels in certain parts of the brain, the modules may as well have a more parallel organization. A diagram outlining some organizational possibilities is presented in Fig. 4.9.

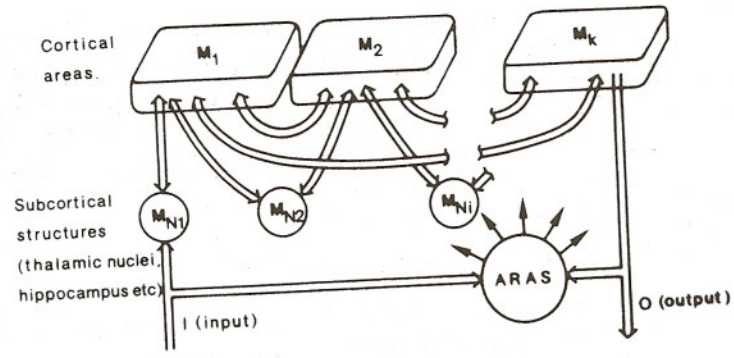

FIG. 4.9. Simplified organization of processing stations, assumed as "matrices," in the brain. The figure is a functional scheme, not showing the true geometry of areas or fibers, the crossing of sensory tracts from the other hemisphere, etc. Mainly those parts are designated that are close to input and output. I = representatives of sensory and other ascending input; O = representatives of descending output to muscles, glands, etc.; ARAS = ascending reticular activation system in the brain stem, which controls the cortical areas.

The interconnected cortical areas are represented in our model by associative memory mappings $\mathbf{M}_i$. It must be understood that these have different degrees of plasticity and memorizing capacity, probably with the plasticity in an adult brain increasing as one moves from the primary sensory areas upwards in the hierarchy. Some or all of the mappings receive afferent inputs of various strengths, although in the schematic figure the inputs have been represented by only one arrow. The afferent input reaches the neocortical areas through thalamic nuclei, depicted in Fig. 4.9 by circles. These contain some mappings $\mathbf{M}_{Ni}$, whose plasticity, if any, is probably much lower than in the cortical areas. Other subcortical units taking part in information processing, like the limbic system, have also been represented by similar circles. The cortical and subcortical units are interconnected in both directions. The efferent cortical outputs are designated by an arrow, again without specifying from which parts of the interconnected net they actually come.

An important subcortical feature in the model is the ARAS, or ascending reticular activating system, located in the brain stem. Its function is to exercise control over the activities in the cortical areas and over the incoming sensory information. It has been empirically shown that the consolidation of memory traces is affected by the activity of ARAS (Bloch, 1976). It thus has the effect of increasing the selectivity and optimizing the resources of the information-processing units by suppressing some activities and enhancing others. This also implies that the plasticity of the network depends on its chemical state or on some other global property.

This description obviously contains a number of gross oversimplifications, which would have to be taken into account in a more detailed model. However it is not the purpose of the present chapter to build a structural model of the brain that precisely fits the available anatomical data but to try to find out what the general ways are in which information processing can be organized using a distributed associative mode of operation.

4.3.3. Synaptic Modification and Learning

Inherent in the laminar network model is the assumption that memory is encoded in the vast number of junction strengths of the interconnections of the network. To be able to write new data into the memory, these junction strengths must be modifiable; to gain selectivity, the modification must depend in some way on the actual signals that are passing through the junction. The convergence of axonal inputs in the cortical lamina ensures that in each column there are synapses signaling the activity of the majority of other columns in the interacting areas. The information needed for cooperative functioning is thus locally available. The selectivity of the memory function is attainable in the network if and only if the synaptic change is limited to some combination of the signal activity. Probably the simplest mechanism proposed to explain the selective synaptic modification is the principle stated by Hebb (1949) in a general nonmathematical form, according to which an individual synaptic junction may increase its efficacy if it is repeatedly activated simultaneously with the triggering of the postsynaptic cell. This hypothesis was presented at a time when the physiological embodiment of learning was thought to be the formation of new stimulus–response connections. Later this hypothesis was expressed as the *conjunction theory* of learning (Eccles, 1978; Marr, 1970): A synapse strengthens if both the presynaptic and the postsynaptic neuron are active at the same time.

One may ask why there has not been a conclusive experimental verification of the existence or nonexistence of conjunction type learning during the thirty years between the appearance of Hebb's book and the present day. The answer lies in the very nature of the conjunction theory. One must realize that this theory does not presuppose that the memory traces are encoded in strong changes in some individual synapses or even in an increasing heterogeneity in the strengths of a number of synapses; the integrated change even in a small population of neurons may be zero due to positive and negative individual changes (cf. Eq. 4–27). The phenomenon is more subtle. It can best be explained by mathematical correlation, which is the macroscopic implication of conjunction learning in individual synapses on the level of the whole network.

It is as impossible to infer the contents of the entire memory from a few junction strengths as it would be to compute, say, the eigenvectors of a large matrix from a few matrix elements. To carry the analogy a bit further, subtle and almost unnoticeable changes in a large number of synaptic strengths can strongly

affect the overall properties of the memory filter just as small selective variations in suitably chosen matrix elements may radically alter the linear mapping whose numerical counterpart is the matrix.

For this reason the prospects of a direct experimental breakthrough in favor of the conjunction hypothesis may not be good. This applies especially to large networks, like those of the mammalian cortex. In small systems of neurons the problem may be easier to solve. In fact the most important advances in this area have been made in the study of small neural systems like that of *Aplysia* (Kandel, 1979). This work has revealed some of the laws governing presynaptic sensitization. In the study of higher organisms most of the results are connected to long-term potentiation phenomena in the hippocampus (Eccles, 1979).

If the conjunction assumption is accepted as a working hypothesis, there are still several possibilities for the actual mechanisms involved. Some of these are growth and regression of presynaptic endings; changes in the transmitter concentrations; and redistribution or activation and passivation of postsynaptic receptors of the cell. These may be accompanied by other more permanent changes in the synapses or in the postsynaptic membrane, in which the memory traces become consolidated.

One of the more detailed and plausible models, although by no means the only one producing conjunction-type learning, is the redistribution of receptors between the synapses of a cell according to demand (Huttunen, 1973; Stent, 1973). This explanation has the advantage of being able to produce very quick changes, which may later be consolidated by a fixation of the receptor molecules onto the cell membrane.

This model has been discussed in detail by one of the authors (Kohonen, 1977). The ensuing equation for synaptic change is

$$\frac{d\mu}{dt} = \alpha\eta(\xi - \xi_0) \tag{4-27}$$

where μ is the efficacy of the synapse, η is the postsynaptic activity given on a frequency scale, ξ is the corresponding presynaptic activity, and ξ_0 is an equilibrium value, the input frequency causing no changes in the value of μ. The scalar α is a constant determined by the level of plasticity of the synapse. The above law applies both to excitatory and to inhibitory synapses and explains both weakening and strengthening, depending on whether ξ is below or above the equilibrium level ξ_0. In fact, the term $(\xi - \xi_0)$ may be considered as the effective presynaptic signal, attaining both positive and negative values (Kohonen, 1977).

The term $\eta(\xi - \xi_0)$, which is the product of postsynaptic and presynaptic signals, gives rise to a correlation matrix appearing in the learning equations, as explained previously in Section 4.2.3. Based on the conjunction form of synaptic plasticity, several typical functional units may be *adaptively* formed depending on the wiring of the network. Some such systems were reviewed in Sections 4.2.3.–4.2.6.

4.4. INFORMATION PROCESSING IN DISTRIBUTED ASSOCIATIVE MEMORY

4.4.1. Sensory Experiences as Patterns of Activity Over Memory Fields

As stated in earlier sections, the brain is an interactive system in which the activity of every part is affected by many other parts. Accordingly the associative mapping that takes place in one hypothetical functional unit can only describe an elementary operation in a complex sequence of processes. It is pointed out in the following discussion that these basic operations may, nonetheless, operate upon semantically meaningful representations of occurrences.

For every part or functional unit in the brain that contains memory we shall use a representation that we henceforth call a *memory field*. (A similar approach was taken by Nakano, 1972.) In its simplest form a memory field corresponds to a lamina of memory elements (cells or tightly connected agglomerates of cells such as columns) that have a great number of mutual connections, assumed to be distributed uniformly over the field (cf. Section 4.3.). The memory field can be identified with the top view of an area of cortex. There is sensory or other primary input to every element in this field. This input is assumed to originate at some prior processing stations (e.g., nuclei or other cortical areas) with the result that different locations in the memory field already have differentiated roles; signals representing particular features are clustered in particular locations. Some locations may be thought to receive input from subcortical structures, so these signals may have an emotional significance.

In the modeling approach it is thus possible to assume that within a memory field certain local areas, which we schematically distinguish by circular regions in Fig. 4.10, have a well-defined modality and meaning. *A region is now identified with an attribute, and the spatial pattern of activity within the region represents the value of that attribute.* (In reality, of course, these areas may be more or less diffuse.) There are good experimentally verified reasons to assume

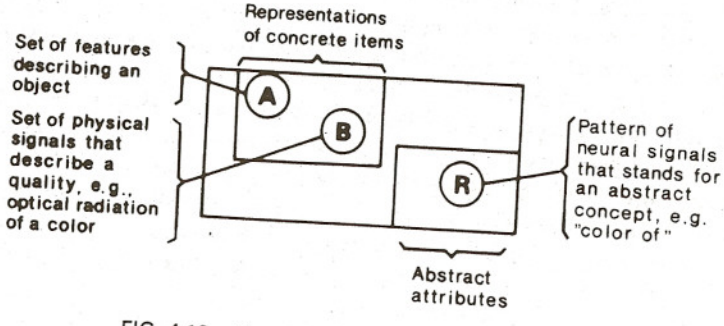

FIG. 4.10. Hypothetical example of a memory field.

that for a particular sensory experience or other occurrence the pattern of activity over the complete memory field consists of only a few activated local areas (attributes) (Barlow, 1972) whereas the rest of the elements in the field are silent, having effective signal value zero (or rather, being regarded as signalless). This theory of representation now makes it possible to compare the memory field model directly with an abstract relational representation.

Assume for simplicity that there are only three active local areas as in Fig. 4.10, labeled A, R, and B. The signals at A and B may stand for features that make up the representations of two items, for example, one being the representation of an object and the other some concrete physical observation, whereas the region for R might contain neural signals that represent an abstract, relational attribute. Notice that the memory field model would allow an arbitrary number of such attributes to be present in one occurrence. It is also important to notice that attributes are bound to a particular location in the memory field.

This model, which uses patterns in memory fields, can now be contrasted with the traditional way of representing relations in data structures by triples of symbolic items, which may be traced back to the use of predicate calculus in early question-answering systems (McCarthy, 1959). If A, B, and R are symbolic (distinct) *items*, then (R, A, B) is a *relational triple:* For example, R = color of; A = an apple; B = is red; (R, A, B) = color of an apple is red. It is possible to question the generality of this formalism in the description of brain functioning. We feel that linguistic expressions have an extremely high degree of coding based on (implicit) assumptions and conventions, like the assignment of a particular meaning to prefixing, suffixing, and other formats. Expressions like (R, A, B) only *look* simple; in fact their treatment in the brain might need complicated processing, probably augmenting the representation of the arguments by contextual information that indicates their role. The concept of a memory field, in which regions are identified with attributes and the values of attributes correspond to spatial patterns within regions, provides a more realistic view of the representation of information in a neural network. The network is then able to process, for example, semantic triples or other tuples of symbolic items, but this is to be regarded as a secondary process rather than its natural mode of functioning.

4.4.2. Generation of Answers to Implicitly Defined Queries Presented to the Laminar Memory Model

It is frequently stipulated that genuine models of memory should be able tc generate answers to complex queries or to perform searches for pieces o memorized information that are only implicitly defined by their relational struc ture. For this reason there has been considerable interest in memory model defined in certain artificial intelligence languages; the answers are found by series of list-structure processing operations. In these, tolerance to errors i

names, labels, and data structures is poor when compared with the performance of biological memory, nor can this type of memory structure generalize over clustered representations without adding considerable extra apparatus to inspect the memory structures.

Because the associative mappings implemented in the "matrix" memories have the ability of representing and retrieving patterned information that may also be clustered, it would be interesting to develop these models in a direction in which they too could be made to search for implicity defined memorized information on the basis of separately given cues. Contrary to what is generally believed, such processes are implementable by rather simple mechanisms, thereby preserving the ability to deal with data that have statistical properties (i.e., not bound to occur in a unique form). It is even possible to demonstrate some elementary forms of thinking and problem-solving processes, although we are still a long way from implementing real thought processes.

As it might be rather difficult to pick up an exemplary system which is general enough and at the same time reflects powerful information-processing abilities, we would like to rest content with a rather simple example that is still demonstrable with the aid of a few illustrations and a little text. First of all we revert to the field representation introduced in Section 4.4.1. Notice that because there exist interactions (associative connections) between all areas of the memory field, the active areas of the field can also be viewed as distinct functional "units." In this way *the operational units can be regarded as virtual ones,* allocated from the memory "field" according to need by an activation and selection system. It is then possible to consider the autonomous computing processes as taking place in a system of these virtual units or "virtual processors," in a series of iterative recall processes.

It seems that a search task in which the target item is implicitly specified by multiple relationships always involves multiple computations that are *separated in time*. Thus, although the memory network from which associative recall is possible is distributed and the operations in it are fully parallel, nonetheless there must exist a phase in the operation in which intermediate results are collected, compared, and collated. It can now be shown that there is no need to devise a complicated processor for this purpose. In many cases the system that carries out this collection and thus assumes the role of a a short-term memory (STM) can be extremely simple. Some sort of retention of the output signals from a functional unit with a duration of, say, several seconds is necessary, and the easiest way to implement this is by some kind of "*leaky integrator,*" (Shiffrin, 1976). We shall not try to specify a physical or chemical mechanism to make this retention possible. Notice, however, that even dynamic reverberation of signals between two units in a point-to-point fashion is a possible STM, and such point-to-point circuits are known to exist between the cortex and the thalamus of a mammalian brain.

In order to collate signal values obtained in matching separate incomplete patterns, the outcomes must first be normalized. This does not mean simply standardization of signal amplitudes to, say, two discrete levels but rather that the *time integrals* of the output signals from the laminar memory ought to be standardized. This is possible if temporal differentiation of the patterns takes place at the input to the matrix itself, or at its output. Again we shall not specify a particular mechanism more closely. One possible effect would be a short-term habituation, but this process might take on rather complex and at the same time intriguing forms (e.g., Kohonen & Oja, 1976).

One further operation in the collating process is carried out by a *threshold trigger*. The threshold might be adjustable by slow adaptation so that the triggering may be made to occur after two or more accumulated output signals have been obtained in successive matches to incomplete search patterns.

An Example of a Searching Process. Consider the data structure shown in Fig. 4.11, which is supposed to represent the contents of a semantic memory; this structure is formed of relational triples (A, P, B), (A, Q, C), etc, in which the middle elements always comprise the *link labels,* and the two others the associated items. Assume that item C had to be retrieved on the basis of two incomplete search patterns (A, Q, X) and (E, R, X) with X unknown. The normal search procedure would first determine the set of solutions for X for each search pattern separately, and then find the value $X = C$ as the intersection of these two sets.

Let us now study how the same task would be solved by distributed memory. The data structure would be stored in terms of triples, which would be represented as patterns defined over the memory field, for example, of the type of Fig.

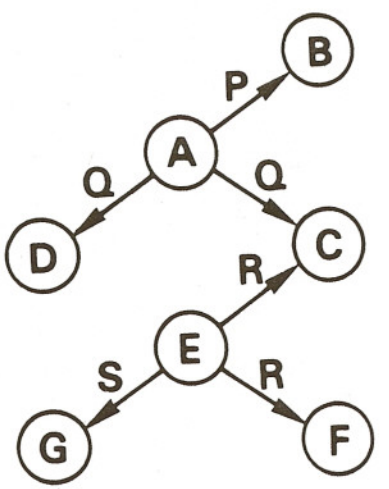

FIG. 4.11. The abstract data structure used in the simulation of Fig. 4.12. Vertices A through G represent items; edges P through S represent link labels.

4.12(a). It is noteworthy that there may be similar local areas in two different patterns, but the areas which are similar will vary from one pair of patterns to the next.

Assume now that two cues, two incomplete key patterns, are given as shown in Fig. 4.12(b) and 4.12(c). If either of these were separately applied as inputs to an autoassociative memory network, the responses from the latter, being mixtures of stored patterns, would be defined by Eq. (4–13) and delineated as in Fig. 4.12(d) and (e). Due to the assumed standardization of output signals, these recollections would remain subliminal. If, on the other hand, the two key patterns were applied one after the other, with a delay that is less than the time constant of the STM, then the component pattern C would be recalled with approximately double the intensity of the other components, and it would therefore exceed the threshold. Figure 4.12(f) displays only the above-threshold signals that now constitute the solution of the search task.

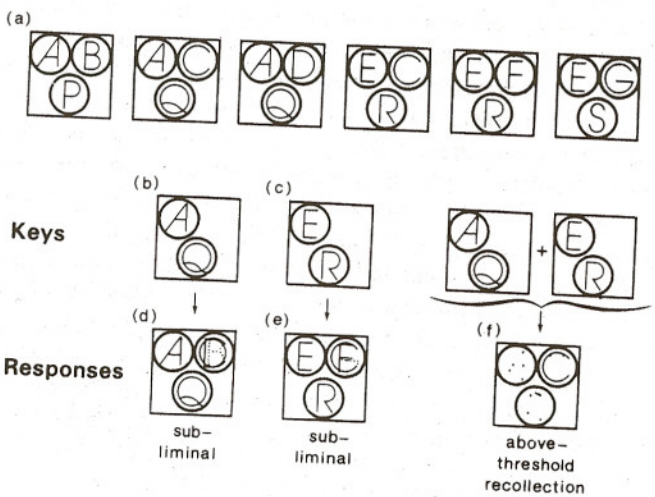

FIG. 4.12. Generation of an answer to an implicitly defined query by the distributed memory model. Part (a) displays in pattern form the relational triples contained in the data structure of Fig. 4.11. The elements of the triples are represented as spatial subpatterns of activity over the memory field. These six patterns have been stored in associative memory. Parts (b) and (c) show two cue patterns applied to the inputs in recall, and (d) and (e), respectively, show subliminal responses that are mixtures of selected images; for example in the upper right-hand subfield of (d), mainly a superposition of C and D is recalled in response to the key pattern (A, Q). When cue patterns (b) and (c) are shown subsequently with a time delay not exceeding the time constant of STM, a superposition of the standardized response patterns (d) and (e) takes place, producing the above-threshold recollection shown in (f). At the upper right-hand subfield of (f), an answer to the query (i.e., C) has thus been produced.

The search problem does not become much more complicated even if significantly more local areas (regions) than three are included in the patterns and if several of them are left unspecified in the keys. Although the internal relations in the patterns would be more complex in this case, nonetheless the phases of associative recall, and the recollection of results by the STM, would take place in essentially the same way as described. On the other hand, such a problem would need much more complicated handling by conventional query languages.

There is an intriguing explanation of the interaction of short-term memory (STM) and long-term (LTM), based on the previous model. As there obviously exist several types of memory effects with different time constants in the nervous system, one might assume that the integrated and processed outcomes from the STM become autoassociatively memorized in this or another network. In a later *recall* process this information can be recalled associatively, etc. This allows rather complicated information structures to accumulate in memory with time, without all of these structures being explicitly present at the same time.

4.4.3. Self-Controlled Operation of the Distributed Memory

Organized computing devices can be built of chips containing only logic gates. Similarly we can imagine an *associative processor,* consisting of many distributed memory units interconnected by bundles of signal lines over which they pass cue patterns and recollections to each other. If the system had closed signal circuits, this passing and transformation of information would proceed recurrently or iteratively with each module adding such information to the patterns as was earlier stored in it. These modules could also standardize their output signal values and extract certain features at their inputs, thus compressing information that was present in the transmitted patterns. Unlike digital computers, such distributed memory systems do not need a highly sophisticated control that opens and closes every signal path in a programmed sequence. If any gating of signal paths exists, it can be of a more or less general nature, comparable to the control of arousal or attention in the nervous system. Coarse spatial control could be achieved by context signals, which activate a subset of units for a particular task. The self-controlled operation of this kind of system should be compared with that of a conventional *analog computer,* especially a *differential analyzer.* The operational units of the latter transform signals internally and pass the results to each other in an asynchronous and highly parallel fashion without external control. The analog computer can be envisioned as an autonomous dynamical system in which computation involves continuous change of the state variables over time.

Although the distributed memory system is assumed to behave in a way grossly similar to an analog computer, there exist some characteristic features that distinguish it from a conventional differential analyzer:

1. The system parameters (the weights) change adaptively due to the occurring signals; that is, *memory traces* are collected that facilitate associative recall.

2. The state variables and the input–output signals between the units are not scalar valued. Processing of information within the units as well as communication between them occurs simultaneously and in parallel over a great number of state variables that constitute the patterns.

3. The outcomes from computations in a memory system are usually not trajectories of signal variables in time as in a differential analyzer. Instead the final state of a memory field can be a stationary pattern, in which case the resulting values of some state variables represent the sought information. Alternatively, the system may run through a sequence of states that represent a dynamic recollection of a memorized occurrence.

4.5. DISCUSSION OF CERTAIN PROBLEMS WHICH ARISE WITH PHYSICAL MEMORY MODELS

4.5.1. The Data-Switching Problem

Conventional information-processing models contain implicit assumptions about computational procedures and underlying programs. Current AI research has most often resorted to procedural models that are directly implementable by present computer hardware and languages developed for it. If these models are also advanced in the context of theoretical psychology, one may easily be misled into assuming that intellectual and especially verbal activities require this type of computation. Very little attention has generally been paid to what digital computation actually means.

In conventional digital computers processing operations are concentrated around central arithmetic–logic circuits and associated registers into which the operands have to be *multiplexed* from the memories. Multiplexing and time-sharing of the computational operations is a natural solution for devices that are based on logic circuits. However it also means extensive data transfers between the operational units during which the format of data must be preserved with high fidelity. Moreover, each particular type of datum must then be guided into a particular destination according to its *role*, which imposes extra requirements on its representation and control. It is highly improbable that this kind of multiplexing or data switching occurs in the neural structures where signals are transmitted at relatively low speeds through more or less fixed pathways and *transformed* during their passage from one processing station to another. It is thus implausible that the representation of information in the central nervous system is based on relational triples where items are ordered by their role.

Instead of assuming that the representations of signals in the central nervous system carry with them some specification of their roles and are thus freely transferable, it would be more natural to assume that *the semantic role of a signal explicitly follows from the particular part of the brain and region in the memory field into which it converges*. This becomes possible because of some degree of genetic predisposition in the gross connectivity, although, as pointed out in Section 4.4, the actual forms of signal patterns within an area may vary with individual experiences.

If brain mechanisms are viewed as highly parallel computing circuits, then ideal parallelism means representation of information that is distributed all over the system. It is then no longer proper to think of information as composed of a great number of more or less independent *records*, a view that has been inherited from serial computers. An occurrence could rather have a representation that occupies the whole system, as the activity patterns over the memory fields do.

The difficulties arising from the data-switching problem with corresponding restrictions set to neural implementation seem to favor memory models that are based on memory fields, that is, the distributed associative memory, as well as the connectionist view.

A particular type of memory field can be made to represent a "connectionist" memory directly. If it is assumed that the *memory elements* of the field (not activity patterns over areas) directly represent items, abstract attributes, etc., then these elements might be named *nodes* as in semantic networks. "Associations" would in this view correspond to direct links between nodes. Obviously this model represents an extreme case in which every single element of the memory field has a semantic interpretation. This is a standpoint adopted by some earlier psychological models of memory (Norman, 1968). Not only must the signals corresponding to every sensory experience then be guided to the nodes with perfect spatial resolution and selectivity, but one must also assume that the location of such a node was determined from the beginning, independently of any sensory experiences.

In fact, in the example of information processing by a laminar network as discussed in Section 4.4, operations were performed by a "hazy connectionist network" in which the local areas corresponded to nodes, and their interactions were mediated through the adaptive lateral connections.

4.5.2. Automatic Formation of Symbols in Associative Mappings

It has been a common view that the central nervous system processes information in symbolic form (Newell & Simon, 1972). This was obviously postulated for the following reasons:

1. Symbols can be made *unique* whereby they facilitate long and complex operation sequences.

2. Symbols can be associated with *concepts,* which in the simplest form are representations of *clusters* in more or less variable occurrences.

3. As the concepts can be defined on different levels of abstraction, a more accurate *meaning* can be given to an occurrence at an increased economy of representation.

An extreme form of symbolism would be a "brain code," the existence of which, however, has never been verified experimentally. Now it has to be emphasized that a symbolic representation need not be identified with a code. It seems sufficient that any substitute pattern that is simpler and more invariant than the original occurrence can be associated with the latter; we have tried to delineate in this paper some possibilities for the embodiment and processing of structured relations between such representations (Sections 4.2.6 and 4.4.2).

A trivial and also common way to define a concept is as supervised association of a symbol to a pattern. However, a much more important and intriguing problem concerns mechanisms by which a symbolic representation could *automatically* be formed in an adaptive system. One possibility, formation of certain discrete states in the system corresponding to distinct statistical "eigenfeatures" of the input information is discussed elsewhere in this book (Anderson & Mozer, Chapter 8, this volume). Sometimes such discrete states are formed by simplification, for example, by amplitude discrimination of signals and dropping of weak segments from sequences. Classification, especially false identification with earlier prototypes, is obviously one process which assigns names to new occurrences.

As to the remaining problem, formation of symbols referring to different levels of abstraction, it is useful to observe that regions in scenes which are more *general* are often also more *constant* and vice versa. This may allow differentiation between hierarchical levels on the basis of conditional probabilities of the subpatterns.

4.5.3. Interdependence of Different Operations in Neural Information Processing

A comprehensive theory of neural information processing has to explain many known phenomena not covered in this chapter. How do we recognize a person on the basis of a scratched old photograph never seen before? How do we allocate attention to different parts of a scene full of objects? What is the basis of the perceptual invariances?

The analytical models in this chapter are the results of the study of basic learning paradigms in the nervous system. They try to explain how the associative storage and recall may be implemented in a neural network. Their ability to explain perceptual invariances or complex mental processes is limited. We feel, however, that even if all cognitive processes are tightly interrelated complex processes, there exist subproblems that may be tackled separately.

The brain, which is the most effective adaptive system, is obviously optimized at all levels, in organization as well as in details. Therefore, because representation of organized perceptions and structured knowledge is a demanding task, it seems reasonable to assume that there exist information-processing stages in the brain that can extract invariant primitives of information from signal patterns. Effective feature extraction and standardization functions are known to exist at least in the primary sensory systems, and they have been subject to many theoretical studies; similar standardizing transformations may be found at higher levels of neural hierarchy, too. At least in theory it then seems profitable to separate the modeling of preprocessing from that of memory functions, which allows a more lucid discussion of the latter.

SUMMARY

The major theme in this chapter has been processing of information in distributed associative memory systems that serve as models of adaptive neural systems. This approach provides an answer to one of the most intriguing problems in neuroscience: How may the neural tissue, which is rather uniform over the cortex, be adaptively specified to carry out the highly differentiated functions found in it.

The chapter was divided into three main parts: We first presented in Section 4.2 some models to explain the basic functioning of the distributed associative memory in system-theoretical terms, using optimal associative mappings. It was shown that items of information can be made to correspond to distributed transformation functions that, although not being stored in spatially separate locations, still preserve the distinctness of representations in the recall process. In Section 4.3 the biological feasibility of these system-theoretical ideas was studied in connection with a laminar network model of the neocortex. In this model information is processed in interconnected cortical areas that make it possible to represent the known specificity of different cortical areas and still apply the principle of distributed associative memory. In Section 4.4 the problem of representing semantic data and generating answers to implicitly defined queries in distributed memories was approached and a solution was outlined in terms of laminar memory fields.

REFERENCES

Albert, A. *Regression and the Moore–Penrose pseudoinverse*. New York: Academic Press, 1972.

Anderson, J. A., & Silverstein, J. W. Reply to Grossberg. *Psychological Review,* 1978, *85,* 597–603.

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

Bellman, R. *Introduction to matrix analysis*. New York: McGraw-Hill, 1960.

Bloch, V. Brain activation and memory consolidation. In M. R. Rosenzweig & E. L. Bennett (Eds.). *Neural mechanisms of learning and memory*. Cambridge, Mass.: MIT Press, 1976.

Bush, R. R., & Mosteller, F. *Stochastic models for learning*. New York: Wiley, 1955.

Creutzfeldt, O. (Ed.) Afferent and intrinsic organization of laminated structures in the brain. *Experimental Brain Research*, 1976, *Supplementum 1*.

Eccles, J. C. An instruction-selection hypothesis of cerebral learning. In P. Buser & A. Buser (Eds.), *Central correlates of conscious experience*. Amsterdam: Elsevier, 1978.

Eccles, J. C. Synaptic plasticity. *Naturwissenschaften*, 1979, *66*, 147-153.

Hebb, D. O. *Organization of behavior*. New York: Wiley, 1949.

Hubel, D. H., & Wiesel, T. N. Sequence regularity and geometry of orientation columns in the monkey striate cortex. *Journal of Comparative Neurology*, 1974, *158*, 267-297.

Huttunen, M. O. General model for the molecular events in synapses during learning. *Perspectives in Biological Medicine*, 1973, *17*, 103-108.

Kandel, E. R. Small systems of neurons. *Scientific American*, 1979, *241*, 60-70.

Kohonen, T. Correlation matrix memories. *IEEE Transactions on Computers*, 1972, *C-21*, 353-359.

Kohonen, T. *Associative memory—A system-theoretical approach*. Berlin: Springer-Verlag, 1977.

Kohonen, T. *Content-addressable memories*. Berlin: Springer-Verlag, 1980.

Kohonen, T., & Oja, E. Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biological Cybernetics*, 1976, *21*, 85-95.

Kohonen, T., Lehtiö, P., Oja, E. Kortekangas, A., & Mäkisara, K. Demonstration of pattern processing properties of the optimal associative mappings. *Proceedings of the International Conference on Cybernetics and Society*, Washington, D.C., Sept. 19-21, 1977, 581-585. (b)

Kohonen, T., Lehtiö, P., Rovamo, J., Hyvärinen, J., Bry, K., & Vainio, L. A principle of neural associative memory. *Neuroscience*, 1977, *2*, 1065-1076. (a)

Lewis, T. O., & Odell, P. L. *Estimation in linear models*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Marr, D. A theory for cerebral cortex. *Proceedings of the Royal Society (London)*, 1970, *B176*, 161-234.

McCarthy, J. Programs with common sense. In D. V. Blake & A. M. Uttley (Eds.), *Proceedings of the Symposium on Mechanization of Thought Processes*, 1959. London: H. M. Stationery Office.

Mountcastle, V. B. Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of Neurophysiology*, 1957, *20*, 408-434.

Nakano, K. Associatron—A model of associative memory. *IEEE Transactions on Systems, Man, and Cybernetics*, 1972, *SMC-2*, 380-388.

Newell, A. Reasoning, problem solving, and decision processes. A paper presented at Attention & Performance VIII Conference, Princeton, 1978.

Newell, A., & Simon, H. A. *Human problem solving*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Norman, D. A. Toward a theory of memory and attention. *Psychological Review*, 1968, *75*, 522-536.

Oja, E. *S*-orthogonal projection operators as asymptotic solutions of a class of matrix differential equations. *SIAM Journal on Mathematical Analysis*, 1978, *9*, 848-854.

Quillian, M. R. Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge, Mass.: MIT Press, 1968.

Rao, C. R., & Mitra, S. K. *Generalized inverse of matrices and its applications*. New York: Wiley, 1971.

Shepherd, G. M. *The synaptic organization of the brain*. New York: Oxford University Press, 1974.

Shiffrin, R. M. Capacity limitations in information processing, attention, and memory. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes, Vol. 4*, New York: Wiley, 1976.

Simon, H. A. The information-storage system called "Human Memory". In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanisms of learning and memory*. Cambridge, Mass.: MIT Press, 1976.

Stent, G. S. A psychological mechanism for Hebb's postulate of learning. *Proceedings of the National Academy of Sciences*, 1973, *70*, 997-1001.

Szentágothai, J. Specificity versus (quasi-) randomness in cortical connectivity. In M. A. B. Brazier & H. Petsche (Eds.), *Architectonics of the cerebral cortex*. New York: Raven, 1978.

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