ORIGINAL CONTRIBUTION

Pattern Recognition by Labeled Graph Matching

Christoph von der Malsburg
Max-Planck-Institut für Biophysikalische Chemie
(Received and accepted 1 December 1987)

Abstract—A model for position invariant pattern recognition is presented. Although not demonstrated here, the system is insensitive to distortions. Recognition is based on labeled graph matching. The system consists of two layers of neurons, an input layer, and a memory and recognition layer. The latter consists of subnets to represent individual patterns. In both layers, patterns are represented by labeled graphs. Nodes are "neurons," labels are local feature types, links are implemented by excitatory connections and represent topology. Recognition is driven by spontaneous dynamic activations of local clusters in the input layer. Network dynamics is able to selectively activate with good reliability corresponding clusters in memory layer. Few cluster activations suffice to identify the subnet and pattern corresponding to the graph in the input layer. The system has been implemented and tested with the help of simulations.

One of the driving forces for the present wave of interest in neural networks is the prospect of practical applications. Classical concepts of computer science and artificial intelligence need further development in the direction of self-organization and massive parallelism. This is what neural networks seem to be offering. The availability of cheap processing power on massively parallel machines will permit the expansion of scale of neural networks from demonstrations of principle to more realistic applications. In this process of expansion a number of difficulties with neural systems will surface. One of these has to do with the great variability of natural scenes which are to be processed as input. If neural systems are to absorb information from one scene and apply it to another, they have to be capable of generalization. Important types of generalization can be based on the decomposition of scenes into standard objects and on object recognition invariant under certain transformations. Speaking of the visual modality, important transformations are translation, dilatation, rotation, and distortion, the latter mainly due to changes in perspective. The solution of these and other problems involved in scene analysis may necessitate new conceptual developments or even modifications to neural architecture. This paper tries to contribute to that goal. It presents a specific kind of neural architecture to solve the problem of invariant object recognition. It is a further development of a model presented earlier (Bienenstock, 1988; von der Malsburg & Bienenstock, 1987).

INVARIANT PATTERN RECOGNITION IN NEURAL ARCHITECTURE

Associative Memory (Anderson, 1970; Cooper, 1974; Hopfield, 1982; Kohonen, 1977; Palm, 1980; Steinbuch, 1961; Willshaw, Buneman, & Longuet-Higgins, 1969) has a number of properties which qualify it as an important neural paradigm. Among these is its ability to generalize over Hamming distance. One aspect of this is pattern completion. Translation invariance, on the other hand, is not among these properties. It is not feasible to solve this problem by exhaustively storing all different transformed representations of a given object. This is extremely uneconomical and it is unclear how this multiplicity of patterns could be stored. Either all transformed versions of the object would have to be shown to the system, a ridiculous proposal, or else all transformed versions somehow would have to be automatically produced within the system after it has seen a single representative.

A way out of the difficulty is the introduction of multilayer structures in which input data are preprocessed. This has first been proposed by Rosenblatt (1961) in the form of his "four-layer perceptron." This

Requests for reprints should be sent to Christoph von der Malsburg, Computer Science Department, University of Southern California, Los Angeles, CA 90089-0782.
consists of layers \( S, L^{(1)} \) and \( L^{(2)} \), connected in sequence, \( S \rightarrow L^{(1)} \rightarrow L^{(2)} \). (The fourth layer, \( R \), is not of interest here.) \( S \) is the sensory surface (retina). Layer \( L^{(1)} \) contains “associator units,” or feature detector cells, nowadays called “hidden units.” Each \( L^{(1)} \)-unit receives connections from a specific subset of the cells of \( S \) and is fired by a specific pattern on this support. Cells in \( L^{(1)} \) carry a label \((\omega, x)\), where \( \omega \) is a feature type and \( x \) is a position within the input plane. If the image is shifted to a new position in \( S \), this shifted image leads to the activation of another set of neurons corresponding to a new distribution of feature types over \( L^{(1)} \). It is assumed that this new distribution is a shifted version of the old distribution (apart from border effects), that is, the same feature types appear in shifted positions.

Layer \( L^{(2)} \) contains cells which are labeled by feature type \( \omega \) only. A cell of type \( \omega \) in \( L^{(2)} \) receives input connections from all cells of the same type in \( L^{(1)} \), irrespective of position. A cell in \( L^{(2)} \) can be fired by a single cell of the same feature type in \( L^{(1)} \). (This version is designed to deal with translation invariance. The idea can be generalized to other types of invariance.) When presenting a pattern \( P \) to \( S \) in position \( x \) a set \( L^{(2)} \) of cells in \( L^{(1)} \) becomes active. This in turn leads to the activation of a set \( L^{(2)} \) in \( L^{(2)} \). Whereas the set \( L^{(1)} \) depends on the position of \( P \), \( L^{(2)} \) does not and is thus a position-invariant representative of the pattern \( P \). Therefore, activity pattern \( L^{(2)} \) in \( L^{(2)} \) can be simply stored by associative memory.

The fundamental difficulty with the four-layer perceptor is ambiguity of the representative \( L^{(2)} \) for pattern \( P \). Different patterns \( P', P'', \ldots \) may lead to the same activity pattern \( L^{(2)} \) in \( L^{(2)} \) if they are composed of the same feature types in a different arrangement. The problem is due to the fact that the four-layer perceptor discards position information on the way from \( L^{(1)} \) to \( L^{(2)} \) (as it is designed to do) but it also forgets topological arrangement!

This difficulty can be solved if a set of feature types can be found which is complete in the sense that a rearrangement of relative positions of some features always leads to the appearance or disappearance of others, such that \( L^{(2)} \) becomes unambiguous. This seems indeed to be possible. Let the sequence \( P(i_1, i_2) \), \( 0 \leq i_1, i_2 \leq N - 1, \) be the input pattern, that is, the light intensity distribution on the \( N \times N \) cells of retina. Let \( L^{(1)}(\omega_1, \omega_2) \) be the discrete Fourier series of the pattern, that is,

\[
L^{(1)}(\omega_1, \omega_2) = \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1} P(i_1, i_2) e^{2\pi i (\omega_1 i_1 + \omega_2 i_2)},
\]

with \( 0 \leq \omega_1, \omega_2 \leq 1, \) \( r = 1, 2 \) the spatial frequencies. These numbers are complex, position information being implicit in their phase. With the transition to \( L^{(2)} \) this information is to be discarded, that is,

\[
L^{(2)}(\omega_1, \omega_2) = |L^{(1)}(\omega_1, \omega_2)|.
\]

According to a mathematical result (Hayes, 1982) \( P \) can almost always (in the mathematical sense) be uniquely recovered from \( L^{(2)}(\omega_1, \omega_2) \) up to inversion and shift (see also Gardenier, McCallum, & Bates, 1986; Sanz, 1985). This result is valid for spaces with two or more dimensions. Apparently it has to do with the multiple connectedness of higher dimensional spaces. Thus, magnitudes of Fourier components form a feature set which is complete in the above sense.

Can this result be generalized to wider transformation spaces than just translation? Generalization to scale and orientation invariance may be feasible (Reitboeck & Altman, 1984). Distortion invariance, however, which is so important to deal with changes in perspective, can with all likelihood not be achieved in a system based on Fourier components. The reason is that Fourier components are globally sensitive to the image and their magnitudes are therefore changed by distortions.

It is likely that the solution to the problem of distortion invariance will have to be based on local feature types, that is, cells which are sensitive to patterns in only a small subregion of retina. This is implemented in a biologically realistic way by Gabor functions (Gabor, 1946). On the other hand, local feature detectors very likely lead to rearrangement ambiguity in \( L^{(2)} \). It is conceivable that this dilemma can be resolved with the help of a multiresolution scheme (Burt & Adelson, 1983; Porat & Zeevi, in press; Rosenfeld, 1984), in which there is a coexistence of feature types of different resolution and size of retinal support. High-resolution-small-support cells encode fine detail in a way which is relatively insensitive to elastic distortions, whereas low-resolution-large-support cells guard against permutation of local regions. Further mathematical work is required to explore the feasibility of such a scheme for invariant pattern recognition and representation.

**DISAMBIGUATION BY REPRESENTING TOPOLOGY**

I would now like to look at the problem from another perspective. Assume \( L^{(1)} \) is composed only of local feature detectors \( L^{(1)}(x, \omega) \) with supports of small diameter in \( S \). Assume that neighborhood relationships within \( S \) of feature detectors are encoded in a way which can be transmitted to \( L^{(2)} \). For the sake of clarity, let \( L^{(1)}(x, \omega) \) and \( L^{(2)}(x', \omega') \) be two active feature cells in \( L^{(1)} \) responding to pattern \( P \) in \( S \); then let \( \delta(x - x') \) be some monotonously decreasing function of the distance of the two features. Assume further that in the transition \( L^{(1)} \rightarrow L^{(2)} \) position information is discarded, as before, but that for each pair of position-independent cells in \( L^{(2)} \) the value of \( \delta \) is known, then the topological arrangement of features in \( L^{(1)} \) and in \( S \) could still be recovered and there would be no ambiguity of rearrangement.
How can such a scheme be implemented in neural architecture? A first logical possibility consists in a homogeneous feature set in which the topological relationship of some features is encoded in other features. The magnitude-of-Fourier-components scheme I just discussed may perhaps be regarded as of this type. A second possibility is a layered structure in which feature types on a higher level encode local combinations of features on a lower level. Such systems are described in Fukushima (1980) and Marko and Giebel (1970). Both of these systems are capable of recognition invariant with respect to translation and distortion. Patterns to be recognized are trained into these systems. It was demonstrated that small sets of patterns (black-and-white characters) could be classified. It is not clear how such systems would scale with the number of patterns stored. It is conceivable, though, that some saturation level can be reached in which the machinery on low levels of the system is complete in the sense that no new feature types are needed for new patterns. It would be desirable to find such complete feature hierarchies and characterize them mathematically.

In the systems just discussed cells in higher levels of the hierarchy represent a mixture of feature types and topological relationships. Alternatively one could keep the two types of information conceptually distinct and represent them differently. The network $L^{(1)}$ would then be able to represent retinal patterns as labeled graphs. Their nodes are neurons, labels are feature types, and their links are neighborhood relationships. If also in $L^{(2)}$ patterns are represented as labeled graphs, the problem of pattern recognition is reduced to that of labeled graph matching. The mathematical problem of subgraph isomorphism in general is computationally intractable (Garey & Johnson, 1979). For restricted classes of graphs, among which there are planar graphs, isomorphism can be tested in polynomial time (Filotti & Mayer, 1980; Miller, 1980). It is not clear how, precisely, graph matching as required for invariant pattern recognition is to be formulated mathematically. The difficulty lies in introducing sloppiness in appropriate form so as to arrive at robust pattern recognition.

Another issue is to find an appropriate neural architecture to represent graphs and to implement graph matching. There are several papers proposing to represent links in a graph again by neurons, such that there are node-neurons and link-neurons (Hopfield & Tank, 1986; Kee & Zippelius; submitted for publication: Phillips, Hancock, Willson, & Smith, 1988). The alternative to this is the representation of links by dynamical connections between node-neurons and by temporal correlations between their signals (Bienenstock, 1988; Bienenstock & von der Malsburg, 1987; von der Malsburg, 1981, 1985; von der Malsburg & Bienenstock, 1986, 1987). The advantage of the latter architecture is its parsimony in terms of network structure. It performs graph matching not because it has been specifically wired in, but because it is a natural process in neural nets. This paper explores a specific version of the dynamical link architecture which is extreme in the sense that it relies entirely on temporal signal correlations to represent links and renounces at rapid modification of synapses.

**THE MODEL**

As in the previous discussion, the model consists of two networks, $L^{(1)}$ and $L^{(2)}$, to represent and recognize input patterns, respectively. Neurons are labeled by feature types, of which there are $N_f$. The structure of $L^{(1)}$ is a two-dimensional lattice of points. Each point of this lattice has a full complement of feature types. It is imagined that when an image is presented to the retina $S$ some low level vision mechanism selects one neuron per node of the grid of $L^{(1)}$. The distribution of feature types over the plane of $L^{(1)}$ represents the image projected into the retina. (This restriction to one cell per node is made for the sake of simplification. It is easy to generalize to several active feature cells per point.) Neural dynamics is regulated such that cellular signals are temporally unstable. Excitatory connections between neighboring cells in $L^{(1)}$ induce correlations between signals of these cells. The function $d$ of our abstract scheme is thus represented by pairwise correlations of signals.

Network $L^{(2)}$ consists of several sub-networks, or "subnets." Also these subnets are structured as two-dimensional lattices of points, similar to $L^{(1)}$, although possibly smaller in extent. Each point in a subnet lattice is occupied by one neuron and thus by one feature type. In this way each subnet of $L^{(2)}$ represents a particular pattern. Recognition now has to take the following form. If an image creates in part of $L^{(1)}$ a distribution of feature types which is an isomorphic replica of the distribution stored in one of the subnets of $L^{(2)}$, then the system is to recognize this pattern by activating the subnet. This is done with the help of a sequence of activations of local clusters of neurons in $L^{(1)}$. Each such cluster is created as result of the exchange of short-range excitation and long-range inhibition within $L^{(1)}$.

A cluster contains a small subset of all possible feature types. Connections from $L^{(1)}$ to $L^{(2)}$ activate those neurons in $L^{(2)}$ which have the same feature types contained in the active cluster in $L^{(1)}$. Among the neurons activated in $L^{(2)}$ there are two kinds. One is a random collection. The other forms a local cluster in that network which is isomorphic to the one in $L^{(1)}$. Exchange of excitation and inhibition within $L^{(2)}$ now switches off the isolated random cells and stabilizes the local cluster. After a short sequence of such paired cluster activations in $L^{(1)}$ and $L^{(2)}$ it is possible to identify the subnet in $L^{(2)}$ which is isomorphic to the active network in $L^{(1)}$ and which recognizes the pattern. Informally, a cluster activation in $L^{(1)}$ amounts to a description of a local
neighborhood of the pattern presented, whereas a cluster activation in $L^{(2)}$ identifies a similar neighborhood in one of the stored patterns.

The formulation given so far is identical to that contained in Bienenstock and von der Malsburg (1987) and von der Malsburg and Bienenstock (1987). In the neural architecture proposed here, each cell $i$ has two types of signals, an "external signal" $\sigma_i$ and an "internal signal" $\varphi_i$. The $\sigma$-signals assume the two values 0 or 1, whereas the $\varphi$-signals range continuously within the interval $[-1, +1]$. Only $\sigma$-signals are communicated over connections from one neuron to others, whereas the $\varphi$-signals are confined to the interior of neurons. There are two types of connections, labeled $T_{ij}$ and $K_{ij}$, $T_{ij} \in \{0, a_T\}$ with $a_T$ a parameter and $T_{ii} = 0$, whereas the $K_{ij}$ may be any real number. The dynamics of $\sigma$-signals is assumed to be described by a Boltzmann distribution, that is, the probability of state $\{\sigma\}$ is proportional to $\exp(-H(\sigma)/T_s)$, with $T_s$ being a "temperature"-parameter and $H(\sigma)$ the energy function

$$H(\sigma) = -\sum_{i,j} T_{ij} \sigma_i \sigma_j + \gamma \sum_i p_i - \sum_i \sigma_i^2 - \sum_i \varphi_i \sigma_i.$$  

If $T_s$ is low, only states with low $H(\sigma)$ are realized. The first term is due to (Hopfield, 1982). It favors states in which neurons coupled by a $T_{ij} > 0$ are in the same state, both 1 or both 0. The second term prefers states in which exactly $p_i$ neurons are on ($\sigma_i = 1$). The third term favors $\sigma_i = 1$ if neuron $i$ has $\varphi_i > 0$, and favors $\sigma_i = 0$ if $\varphi_i < 0$. The dynamics of the $\sigma$-signals is illustrated here by the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1983), such that the internal signals $\varphi_i$ are considered fixed during variation of $\sigma$-signals.

The dynamics of the internal signals with given $\{\sigma\}$ is formulated by the energy function

$$H(\varphi) = -\sum_i K_{ij} \varphi_i \sigma_j.$$  

Since $\varphi$-signals are continuous-valued, their dynamics are simulated as gradient-descent, that is, for a given $\sigma$-configuration, $\varphi_i$ is modified by the replacement

$$\varphi_i \rightarrow \varphi_i + \beta_{\varphi} \sum_j K_{ij} \sigma_j,$$  

where $\beta_{\varphi}$ is a parameter. The composite dynamics of both types of signals is simulated by alternating a number of attempted flips of $\sigma$-signals and one step of modification of $\varphi$-signals. It is assumed that parameter $\beta_{\varphi}$ is small so that $\sigma$-signals have time to fluctuate through several configurations before significant modifications of $\varphi$-signals take place. For the functional and biological motivation and justification of this two-tiered version of neural architecture see below.

Network $L^{(1)}$ has only $T$-connections, no $K$-connections. Pairs of neurons occupying neighboring lattice points are connected ($T_{ij} > 0$), irrespective of feature type. All other $T$-connections within $L^{(1)}$ are zero. Networks $L^{(1)}$ and $L^{(2)}$ are connected by feature type specific $T$-connections: a neuron in $L^{(1)}$ is connected to a neuron in $L^{(2)}$ if they carry the same feature type label. Within subnets of $L^{(2)}$ neurons occupying neighboring lattice points are connected by $T$-connections, irrespective of feature type. A neuron in $L^{(2)}$ may belong to several subnets, in which case the set of its connections is the union of all connections required for the subnets it belongs to. Define

$$\xi^a = \begin{cases} 
1, & \text{if } i \text{ belongs to subnet } \alpha \text{ of } L^{(2)}, \\
-1, & \text{otherwise.} 
\end{cases}$$

Then the $K$-connections within $L^{(2)}$ are defined as

$$K_{ij} = \frac{1}{2N} \sum \xi^a (\xi^a + 1).$$

Here, $N$ is the total number of cells in $L^{(2)}$.

**SIMULATIONS**

The simulations described here have been designed with the sole intention of demonstrating an idea, and they are therefore of utmost simplicity. System generation starts with the definition of a small number of patterns, which are stored in $L^{(2)}$, and which are later to be displayed in $L^{(1)}$ and recognized in $L^{(2)}$. Each pattern is a 10 × 10 square grid of feature types. These are selected randomly from the full set of size $N_f$. Typically, $N_f = 50$. These patterns are stored in $L^{(2)}$ by an algorithm (pattern storage by self-organization is not attempted here) which takes one of two possible forms, depending on whether the subnets of $L^{(2)}$ are to overlap or not. In the case of disjoint subnets, individual neurons are simply assigned to the lattice positions of a pattern. With positions in a pattern, neurons acquire feature types. After neurons are positioned within a pattern, they are $T$-connected with their four direct neighbors (border cells have fewer connections). Since subnets are to be disjoint, a new set of 100 cells is selected for each new pattern. If subnets are allowed neuron-wise overlap, neurons are picked randomly from a set of $N_{stock}$ neurons available in $L^{(2)}$. A neuron can only be selected once for one pattern. Furthermore, if a cell is assigned to positions in several patterns, these positions have to have the same feature type. To this end, $N_{stock}$ is divided into blocks of equal size, each block corresponding to one feature type, and neurons are selected randomly from the correct block according to required type. In the end neurons are renumbered in the sequential order in which they have been selected. See Figure 1 for an example with three patterns, constructed with $N_{stock} = 500$.

An actual run is started by initializing all $\varphi$-signals to the value 0. Then, one of the stored patterns is loaded into an arbitrary region of size $10 \times 10$ in $L^{(1)}$, assigning feature types to positions in this lattice. (Only the rel-
evant region in $L^{(1)}$ is represented in the simulations. Patterns deviating from the stored ones have not been tried.) Now the following sequence of events is repeatedly stepped through: (a) A local block of neurons is activated in $L^{(1)}$. This activation simulates a spontaneous event in the $\sigma$-dynamics of $L^{(1)}$, which has a tendency to form such blocks, see below. The feature types of the neurons active in this block form a set which is referred to as the “active set” of feature types; (b) All those neurons in $L^{(2)}$ whose feature types are not in the active set are clamped to signal $\sigma = 0$. Figure 2, upper row of panels, shows all neurons which are free to fluctuate. Their signal is set to 1; (c) All neurons of $L^{(2)}$ which are not clamped are allowed to fluctuate according to $L^{(2)}$-dynamics. Simulated annealing with energy function (1) is performed, with $4 \times N$ flipping attempts ($N$ is the number of cells in $L^{(2)}$) at non-clamped neurons, at the “temperatures” $T = .3, .05$. A typical result for the case of overlapping subnets in $L^{(2)}$ is shown in Figure 2, lower row. With non-overlapping subnets, dynamics in $L^{(2)}$ is very often able to just activate the correct cluster in the correct subnet; and (d) With the $\sigma$-signals thus obtained $\varphi$-signals are modified through one step.

The function of the system is simple to explain. The active set of feature types present in a local cluster of neurons in $L^{(1)}$ conveys no direct information on position in the image plane since only feature type information, no position information, is sent to $L^{(2)}$. If the feature type distribution in the relevant part of $L^{(1)}$ is identical to that of a pattern and subnet stored in $L^{(2)}$, then there is in that subnet a local block of neurons whose feature types are in the active list. This is clearly seen in Figure 2, upper left panel. In addition there are other neurons, scattered over the subnets of $L^{(2)}$, whose feature types happen to be part of the active list. Now, the task of $\sigma$-signal dynamics in $L^{(2)}$ is to stabilize the correct target block and to deactivate all scattered neu-
The dynamics of $L^{(2)}$ according to (1) favors blocks of a size which is controlled by $p_x$. If this is set appropriately, blocks of the same size as those in $L^{(1)}$ are produced in $L^{(2)}$. The scattered neurons are in a bad position because none or few of the neighbors to which they are $T$-connected are legalized by the active list. Only in rare accidental cases spurious clusters are formed. In the case of overlapping subnets it is of course unavoidable that some cells are active in wrong subnets, see lower row of panels in Figure 2.

Since the $\sigma$-signals react slowly to $\varphi$-signals, according to (2), they are sensitive to the union of a sequence of individual $\sigma$-activations. This union has a smaller Hamming distance to the correct subnet, which in each part-activation had most active neurons, than to the other subnets. As a consequence $\varphi$-signals grow for the neurons in the correct subnet and fall for neurons which are not in this subnet. According to the last term in (1) this leads to a further decrease of the probability of non-members of the correct subnet to $\alpha$-fire, and a further increase for members. The result of the recognition process can be read off either from $\alpha$-signals or from $\varphi$-probabilities. In the case with no overlap between subnets in $L^{(2)}$ the correlations $C_{\alpha} = \sum \varphi \alpha \varphi \alpha \alpha \alpha$ between $\varphi$-signals and membership function (3) after 13 cluster activations analogous to the one in Figure 2 reached the values $C_1 = 1.0$ and $C_2 = C_3 = -\frac{1}{4}$. This is the optimum obtainable, since $-\frac{1}{4}$ is the correlation between different patterns. In the case with overlap between subnets, 20 cluster activations sufficed to reach optimal correlations. In another simulation, in which subnets were generated with $N_{\text{stock}} = 250$, which led to overlaps of 43, 42, and 44 cells between subnets 1 and 2, 1 and 3, and 2 and 3, respectively, discrimination was still possible with 20 cluster activations ($C_1 = .691$, $C_2 = .116$, $C_3 = .065$), but the optimal values could not be reached, and longer runs even let correlations degrade.

DISCUSSION

The mechanism of graph matching implicit in the model presented here relies on labels on nodes. Graph matching can be done very efficiently if individual labels are unambiguous. This cannot be achieved with labels in the form of low-level feature types. However, it may be expected that local clusters of features, which collectively describe a whole region in a pattern, may serve as relatively unambiguous labels. Practical experience with real images will have to show the level of reliability which can be reached with individual cluster activations.

No attempt has been made here to exploit fast synaptic modifications used in other formulations (Bienenstock & von der Malsburg, 1987; von der Malsburg, 1985; von der Malsburg & Bienenstock, 1986, 1987). Their function has partly been replaced by $\alpha$-signals. This leads to a great speed-up of the recognition process, since a subnet in $L^{(2)}$ needs not to be covered completely by active clusters in order to be activated. This system
is, however, limited to recognition of rigidly stored patterns. Dynamical links are required for object and scene analysis, see below. Dynamical links were required in von der Malsburg (1985), and von der Malsburg and Bienenstock (1986, 1987) because in those formulations the much harder situation was treated in which different subnets overlap in all their cells and differ only in their connectivity.

In Bienenstock and von der Malsburg (1987) it had been tried to give a formulation of extreme simplicity. Complete overlap of subnets in \( L^{(2)} \) precluded the introduction of an equivalent to \( \theta \)-signals. Only two feature types, black and white, were used, and dynamics was based on 4-cycles of links, connecting pairs of neighboring cells in \( L^{(1)} \) to pairs of neighboring cells in a subnet of \( L^{(2)} \), for instance. This corresponds to clusters of only two cells in \( L^{(1)} \) and in \( L^{(2)} \). Its simplicity of formulation gives the model (Bienenstock & von der Malsburg, 1987) a chance of analytical treatment (see Werman, 1987), but makes it impractical. The model presented here stresses the extreme of high numbers of feature types, low overlap between subnets in \( L^{(2)} \) and large activity clusters, all of which reduces ambiguity of matching and reduces computing time.

Before the model is practically applied it has to be generalized with respect to lattice type and number of active cells in each grid point of \( L^{(1)} \) and \( L^{(2)} \). As many features as possible should be extracted from images in order to obtain maximum reliability in finding matching clusters in stored subnets. However, feature classification must be robust against noise and against transformations over which the system is to generalize. For instance, if edge orientation is to be a feature type then the system cannot generalize over orientation. Provided that features are position invariant, as assumed in the introduction, and size invariant within a certain band, the system is capable of complete translation invariance, limited size invariance and distortion invariance.

The problem of how to store new subnets in \( L^{(2)} \) has not been treated here. In principle it is easy. The system finds out that a new pattern is presented by failing to recognize it. It then has to go into a different mode, in which it selects one cell from \( L^{(2)} \) (randomly or otherwise) for each active cell in \( L^{(1)} \). It then ties together the selected cells in \( L^{(2)} \), globally by \( K \)-connection, in a way described by (4) which is analogous to associative memory, and topologically by \( T \)-connections. For the latter the rule is that if two cells in \( L^{(1)} \) are connected then their images in \( L^{(2)} \) are to be connected also. The difficulty with this scheme is that the selection of new image cells in \( L^{(2)} \) is arbitrary and therefore cannot be easily made in an incremental way over several presentations of a new pattern. It therefore seems more realistic to require that new networks be created in \( L^{(2)} \) by combining already existing subnets which correspond to familiar subpatterns in a new object encountered. This would require dynamic connections within \( L^{(2)} \).

It is not intended to imply that recognition of complicated objects is to be done with a two-layer network, consisting just of \( L^{(1)} \) and \( L^{(2)} \). There are great advantages in having a cascade of such layers. At lower levels the system would recognize more general, abstract object classes. The result of this is the activation of object-specific cells. These could, on a higher level, be interpreted as feature detectors of a much more specific kind than those at lower level. Complicated objects would, in a multilayer system, be represented by the activation and cross-indexing of many subnets referring to different structural aspects. Such a system would go far beyond the mere classification of objects and would represent objects in the same way as whole scenes. The binding of subnets to each other would require dynamic links.

Is the model biologically relevant? The \( \sigma \)-signals correspond to axonal action potentials. Cellular activity as recorded from cortical cells has rich temporal structure. Its interpretation in terms of local cluster-activations poses much less of a problem than the more conventional assumption of stationary activity levels over periods of a few hundred milliseconds. The intraneuronal \( \varphi \)-signals have to have a much slower time course than \( \sigma \)-signals, influence the probability of cell firing, and are influenced themselves by \( \sigma \)-signals over \( K \)-connections. It is conceivable that \( \varphi \)-signals can be identified with depolarizations and hyperpolarizations in distal dendritic branches, the effect of which is felt near the soma as slow waves of dendritic current. A more likely interpretation of \( \varphi \)-signals is their identification with the intracellular concentration of some second messenger which is influenced by a specific type of synapse and which in turn influences membrane or synaptic properties such as to modify firing probability. The existence of two types of connections poses even less difficulty of biological likelihood. If \( \varphi \)-signals are just distal dendritic signals, then the only difference between the two types of connections would be that \( K \)-connections contact cells distally whereas \( T \)-connections contact them near the soma. If \( \varphi \)-signals are chemical concentrations then two types of synapses have to be invoked.

**REFERENCES**


