Transform Invariant Recognition by Association in a Recurrent Network

Néstor Parga *and Edmund Rolls
Oxford University
Department of Experimental Psychology
South Parks Road
Oxford OX1 3UD, England

Abstract

Objects can be recognised independently of the view they present, of their position on the retina, or their scale. It has been suggested that one basic mechanism that makes this possible is a memory effect, or a trace, that allows associations to be made between consecutive views of one object. In this work we explore the possibility that this memory trace is provided by sustained activity of neurons in layers of the visual pathway produced by an extensive recurrent connectivity. We describe a model that contains this high recurrent connectivity and synaptic efficacies built with contributions from associations between pairs of views that is simple enough to be treated analytically. The main result is that there is a change of behavior

*Permanent address: Departamento de Física Teórica C-XI, Ciudad Universitaria de Cantoblanco, Universidad Autónoma de Madrid, 28049 Madrid, Spain

as the strength of the association between views of the same object, relative to the association within each view of an object, increases. When its value is small, sustained activity in the network is produced by the views themselves. As it increases above a threshold value, the network always reaches a particular state (which represents the object) independently of the particular view that was seen as a stimulus. In this regime the network can still store an extensive number of objects each defined by a finite (although it can be large) number of views.

P.A.C.S. 87.30 Biophysics of neurophysiological processes

1 Introduction

Single neurons with responses which are relatively invariant with respect to for example the position, size and even view of an object or face are present in the primate temporal visual cortical areas (see e.g. Gross et al, 1985; Tanaka et al, 1990; Rolls, 1984, 1992, 1994, 1995, 1996b; Rolls, Booth and Treves, 1996). How could such invariant representations be formed? One suggestion is that there is a short term memory trace built into the learning rule implemented in the visual system which enables for example successive views of the same object to be associated together (Foldiak, 1991; Rolls, 1992, 1994, 1995, 1996b). Because the statistics with which objects are normally viewed in the visual world results in different aspects (e.g. views) of the same object being seen close together in time, such a learning rule might enable different views of objects to be associated together. In contrast, views of different objects rather rarely on average occur close together in time. Thus a neuron could learn by a simple form of Hebbian associativity coupled with a short term memory trace to respond to any views of an object, but to no views of other objects. Foldiak (1991) showed translation invariance learning over a one-dimensional input array to be possible for a simple winner-take-all network with a decaying trace of previous neuronal activity and an associative Hebb rule. Rolls (1992, 1994, 1995, see Rolls and Treves, 1997) suggested that invariant representations could be formed in the visual system for two-dimensional images using a multistage architecture, with convergence onto a neuron at any one layer from a small region of the preceding layer, a short-term trace of preceding neuronal activity, soft competition between the neurons to produce distributed representations, and a Hebb-like learning rule. In the theory as suggested, the trace was implemented by a short term trace in the postsynaptic neuron, and this enabled neurons to learn which inputs from the preceding stage tended to occur close together in time.

Possible neurophysiological mechanisms suggested for the trace included the continuing firing of single neurons in the visual system, which often lasts for 300 ms following a 20 ms presentation of a stimulus (Rolls and Tovee, 1994; Rolls, Tovee, Purcell et al, 1994), which could be implemented by recurrent collateral connections between pyramidal cells in the same layer or in adjacent cortical stages in the cortico-cortical hierarchy of stages in the visual system (see Rolls, 1992, 1994, 1995); the rather slow unbinding of glutamate from the NMDA receptors after they have been activated (this may be seen after even 100 ms); and slow changes intracellularly induced after the Ca2+ entry, which is one step in the induction of long-term potentiation. Wallis, Rolls and Foldiak (1993) and Wallis and Rolls (1996) produced a simulation, VisNet, of this theory of the formation of invariant representations in the visual system proposed by Rolls (1992-1996b). The simulation showed that translation, size, and view invariance could be learned by such a network.

In developing these ideas further, we turn in this paper to an approach which allows an analytic formalism to be brought to bear on the issue of the storage capacity of a *recurrent* network which performs for example view invariant recognition of objects by associating together different views of the same object which tend to occur close together in time. The architecture with which the invariance is computed is a little different to that described earlier. In the model of Rolls (1992, 1994, 1996b, and Wallis and Rolls, 1996), the postsynaptic memory trace enabled different afferents from the preceding stage to modify their synapses onto the same postsynaptic neuron (see Fig. 1). In that model there were no recurrent connections between the neurons, although such connections were one way in which it was postulated the memory trace might be implemented, by simply keeping the representation of one view or aspect active until the next view appeared. Then an association would occur between representations that were active close together in time (within e.g. 100-300 ms).

In the present model, there is a set of inputs with fixed synaptic weights to a network. The network itself is a recurrent network. For the purposes of this paper we are concerned primarily with how the recurrent network would operate once the synaptic matrix has been formed, not with how the synaptic matrix is formed. In the discussion section we will describe a set of neuronal operations that could lead to our synaptic matrix. Let us say for the moment that in the case of a recurrent network we expect that the trace rule is implemented on the recurrent collaterals (see Fig. 2).

In the context of recurrent networks we can consider two main approaches. First, one could consider a very simple approach to store in a synaptic weight matrix the s views of an object. This consists in equally associating all the views to each other, including the association of each view with itself (that is, the diagonal terms of the association matrix). Choosing in Fig. 3 an example such that objects are defined in terms of five different views, this might produce (if each view produced firing of one neuron at a rate of 1) a block of 5×5 pairs of views contributing to the synaptic efficacies each with value 1. Object 2 might produce another block of synapses of value 1 further along the diagonal, and symmetric about it. Each object or memory could then be thought of as a single attractor with a distributed representation involving five elements (each element representing a different view). Then the capacity of the system in terms of the number P_o of objects that can be stored is just the number of separate attractors which can be stored in the network. For random fully distributed patterns this is as shown numerically by Hopfield (1982)

$$P_o = 0.14\, C \tag{1.1}$$

where there are C inputs per neuron (and $N = C$ neurons if the network is fully connected). Now the synaptic matrix envisaged here does not consist of random fully distributed binary elements, but instead we will assume has a sparseness $w = \frac{s}{N} n$, where s is the number of views stored for each object, from any of which the whole representation of the object must be recognised. (n is the average number of active neurons in the element, or portion of the pattern, that describes one view). In this case, one can show (as in Gardner 1988, Tsodyks and Feigel'man 1988) that the number of objects

$$P_o = \frac{k\, C}{w\, ln(1/w)} \tag{1.2}$$

where C is the number of synapses on each neuron devoted to the recurrent collaterals from other neurons in the network, and k is a factor of order one that depends weakly on the detailed structure of the rate distribution, on the connectivity pattern, etc. A problem with this proposal is that as the number of views per object increases to a large number (e.g. > 20), the network will fail to retrieve correctly the internal representation of the object starting from any one view (which is only a fraction $1/s$ of the length of the stored pattern that represents an object). This is because if the cue given by one of the views becomes too small, the internal representation of the stimulus will fall outside the attraction basin of the attractor associated to the object.

The second approach, that we will discuss in detail here, is to consider the operation of the network when the associations between pairs of views can be described by a matrix that has the general form shown in Fig. 4. Such an association matrix might be produced by different views of an object appearing after a given view with equal probability, and synaptic modification occurring of the view with itself (giving rise to the diagonal term), and of any one view with that which immediately follows it. [1] The same matrix might be produced not only by pairwise association of successive views because the association rule allows for associations over the short time scale of e.g. 100-200 ms, but might also be produced if the synaptic trace had an exponentially decaying form over several hundred ms, allowing associations with decaying strength between views separated by one or more intervening views. The existence of a regime, for values of the coupling parameter between pairs of views in a finite interval, such that the presentation of any of the views of one object leads to the same attractor regardless of the particular view chosen as a cue, is the main issue that is dealt with in this paper. A related problem also dealt with in this work is the issue of the capacity of this type of synaptic matrix: how many objects can be stored and retrieved correctly in a view invariant way ? As we will show in the model presented here their number grows linearly with the number of neurons. Some of the groundwork for this approach was laid by the work of Amit and collaborators (Griniasty, Tsodyks and Amit 1993).

A variant of the second approach is to consider that the remaining entries in the matrix shown in Fig. 4 all have a small value. This would be produced by the fact that sometimes a view of one object would be followed by a view of a different object, when for example a

[1] Strictly speaking the matrices in Figs. 3 and 4 do not refer to the same thing. In the first case it denotes the matrix of synaptic efficacies while in the second it is the association between two patterns (the matrix X in the next equation (2.3)).

large saccade was made, with no explicit resetting of the trace. On average, any one object would follow another rarely, and so the case is considered when all the remaining associations between pairs of views have a low value.

2 The Model

We will consider a simple model to see that it can store a large number of objects in a view invariant way. In order to have a problem solvable by standard statistical physics techniques (Amit 1988) we will make several simplifying hypothesis. The N neurons in the model are binary objects. Denoting the state of neuron i by S_i, this variable can only take the values $+1$ and -1.

The views of the objects will be chosen as P special states of the network. Different views, even if they belong to the same object, will be taken as uncorrelated random variables. The interpretation of these states is that they are the internal representations in some internal layer of stimuli which might differ in the way an object is presented: different illumination or color, different view or scale. In this internal layer all these possibilities appear as uncorrelated states of the network, and the only relation between these states when they represent the same object is that they will be associated to each other. Notice that we will refer to all these internal states as "views", regardless of the real difference in the stimuli they represent.

We will also assume that the network is fully connected and we will take a coding rate (or sparseness) $w = 0.5$.

By convention we will label the views placing first all those associated to, say, object one, then all the views defining object 2, and so on. We will denote the internal representation of the ν^{th} view by ξ^ν, and the state of neuron j in this pattern by ξ_j^ν $(j = 1, \ldots, N)$ (see Fig. 4). For the sake of simplicity we will assume that all the objects are defined by the same number s of views, in this way the patterns labeled by $\nu = 1, \ldots, s$ refer to views of the first object, etc. As explained above the variables ξ_j^ν will be selected to be $+1$ or -1 with probability $w = 0.5$.

As we said before all pairs of views of the same object are coupled to each other. We will denote the coupling between the patterns ξ^μ and ξ^ν as $X_{\mu\nu}$. Since we are not considering the effect of an association between views of different objects this matrix is made of blocks of size $s \times s$, one block for each of the P_o objects.

These considerations lead us to propose the following form of the synaptic matrix

$$J_{ij} = \frac{1}{2N} \sum_{\mu=1}^{s\,P_o} \xi_i^\mu X_{\mu\nu} \xi_j^\nu \tag{2.3}$$

that, more explicitly, contains a contribution from association between views of the form

$$\frac{1}{N} \sum_{\mu \neq \nu}^{s\,P_o} b_{\mu\nu} \xi_i^\mu \xi_j^\nu . \tag{2.4}$$

where $b_{\mu\nu}$ (the off diagonal elements of $X_{\mu\nu}$) is the strength of the association between views ξ^μ and ξ^ν. The value of the coupling of one view with itself will be taken equal to one. In the case where all pairs of different views of a given object are equally correlated with strength $b_{\mu\nu} = b$ for all $\mu \neq \nu$ while views of different objects are not correlated, the matrix X takes the block form shown in Fig. 4 for $s = 5$. This is the case we will deal with, and the block matrix of size $s \times s$ will be denoted by $\mathcal{O}^{(s)}$. It is natural to think that b is smaller than the strength of the association of one view with itself, i.e. $b < 1$. [2] For an arbitrary value of the sparseness the synaptic matrix given in eq.(2.3) is not a good choice. One should use, for example, the one studied by Tsodyks and Feigel'man (1988). An analysis of the capacity properties for other values of w requires to find the explicit solution of the model for the new synaptic matrix.

Our purpose is to show that this network has an object phase, that is, a phase where each object is represented by a different attractor of the recurrent network. If all views of the same object are contained in its associated attraction basin then view invariant recognition of the object will be achieved when either one of the views or a pattern sufficiently close to one of them is presented as a stimulus.

If each object were defined by a single view, the model above reduces to the Hopfield model which is known to be able to store a number of patterns linear in N. More precisely, in this saturation regime the number of stored objects grows as αN where α has to be less than $\alpha_c = 0.14$. In this case the view is very close to the attractor itself. If this model is modified to contain a finite number of more complex objects (those having several views) in the background of those simple one view objects, it is reasonable to expect that the complex objects will still define stable states if the coupling between the views is strong enough. The case where all the $P_o \simeq O(N)$ objects are defined by a finite number of views requires a more careful analysis. This can be done using tools from statistical physics that have been used some time ago to solve the standard Hopfield model (Amit, Gutfreund and Sompolinsky 1985, Amit 1988).

[2] The choice of equal strength for all the pairs simplifies the numerical analysis; the equations in this section can be easily extended to the general case.

Before presenting the result of this calculation, let us see how to characterize the behaviour of the network for given values of the parameters α and b. The useful quantities for this purpose are the overlaps of the state of the network $\{S_i\}$ $(i = 1, \ldots, N)$ with the views of one of the objects $(\mu = 1, \ldots, s\ )$:

$$m_\mu = \frac{1}{N} \sum_{i=1}^{N} S_i \xi_i^\mu. \tag{2.5}$$

If the network is presented with a stimulus close to one of the views, say ξ^{μ_0}, then all these overlaps will be initially very small except m_{μ_0} that will be close to 1.0. As time goes on these overlaps will change until a fixed point is reached. If the final state is similar to the initial one, then we will say that the network is in the *view phase*. But it could also happen that the stable state has overlaps with more than one view and when it has similar overlaps with all the patterns representing the s views of the object seen initially we will say that this point is in the view-invariant object phase or simply in the *object phase*: because of the symmetry between the views no matter which view is taken as the initial condition of the network that it will always reach the same attractor. Another possibility is that the system will get to a state where all the overlaps are null. Then there is neither view nor object retrieval, and the system is beyond its capacity. The region where this happens is called, by analogy with a similar situation in statistical physics models, a *spin-glass phase*. For the simple model we are proposing here there are no more complex solutions because of the symmetry of the synaptic matrix.

Let us continue by defining the free energy of a system with Hamiltonian

$$H = \sum_{i \neq j}^{N} J_{ij} S_i S_j. \tag{2.6}$$

The free energy is then computed as:

$$f = -\frac{1}{\beta} \lim_{N \to \infty} \frac{1}{N} \ll \log Tr_S \exp[-\beta H] \gg_{\xi}, \tag{2.7}$$

where the angular brackets denote the average over the variables ξ and the symbol Tr_S refers to a sum over all possible states of the network. An extra parameter, β, has been introduced that acts as the inverse of a temperature. However at the end we will consider only the zero temperature limit. This means that the spike emission dynamics is given simply by

$$S_i(t+1) = sign[\sum_{j \neq i}^{N} J_{ij} S_j(t)]. \tag{2.8}$$

In the thermodynamic description given by the free energy f the state of the system is described in terms of *macroscopic* quantities. This means that instead of the overlaps defined in eqn. (2.5), where the microstate $\{S_i\}_{i=1,\ldots,N}$ is used, one should use their thermal averages. These are computed by averaging over all the microstates weighed with the Boltzmann distribution that appears in eqn. (2.7). From now on we will use m_μ to denote these overlaps. More explicitly, denoting the thermal average with single angular brackets, we have

$$m_\mu = \frac{1}{N} \sum_{i=1}^{N} < S_i > \xi_i^\mu. \tag{2.9}$$

To check the existence of a transition from the view to the object phase one has to look for states such that their overlap with all the views of a given object are, in principle, $m_\mu \sim O(1)$

while the overlaps with the views of other objects are zero (actually $O(1/\sqrt{N})$). In terms of these macroscopic quantities, the view phase is characterized by $m_\mu = 0$ for $\mu \neq \mu_0$. On the other hand, in the object phase these s overlaps are non-zero and equal. It turns out that these order parameters are not enough to describe properly the behaviour of the network. This is because there are $O(N)$ overlaps $O(1/\sqrt{N})$ with the views of the other objects. This effect is taken into account by an order parameter, r, that gives the mean square overlap of all the views of the other objects

$$r = \frac{1}{\alpha} \sum_{\mu > s} \ll m_\mu^2 \gg_\xi . \tag{2.10}$$

Finally one also has to consider the possibility that the m_μ's are zero but the order parameter defined as

$$q = \frac{1}{N} \ll \sum_{i=1}^{N} < S_i >^2 \gg_\xi . \tag{2.11}$$

is not zero.

Another relevant point is that the mean-field solution of this model is exact. This is because the network is fully connected. Then the order parameters satisfy a set of mean-field, self-consistent equations. In order to find them one has first to compute and express the free energy in terms of the order parameters and then to extremize it with respect to them.

The evaluation of the free energy of this kind of system is by now standard (Mezard, Parisi and Virasoro, 1987). The free energy and the mean-field equations have been obtained for a general coupling matrix X (although with a different motivation) by Cugliandolo and Tsodyks (1994). We refer the reader to the Appendix I of their work for details of the calculation. Here we just describe the main steps of the algebra and present the results for the free energy and for the mean-field equations of our problem (that is for the case where X decomposes in P_o $s \times s$ blocks as in Fig. 4).

Very briefly, the computation proceeds as follows: One first uses the representation of the logarithm $\log z = \lim_{n \to 0} \frac{z^n - 1}{n}$ in eqn. (2.7). Then each of the n factors in f is identified with a "replica" of the network. Notice that all of them share the same set of internal representations of the views. At this point the quenched average over these variables can be done easily. Now the free energy can be written in terms of integrals over the order parameters q, r and m_μ (only the $O(1)$ overlaps, i.e. $\mu \leq s$). The final step is to make the ansatz that the problem is symmetric under permutations of the n replicas. Then the remaining integrals can be readily solved in the large N limit (keeping α fixed) by the saddle point method.

After some lengthy but straightforward algebra, one obtains that the free energy at zero "temperature" (i.e. for $\beta^{-1} = 0$) is given by:

$$f = \frac{1}{2} \left(\alpha + \sum_{\gamma,\lambda=1}^{s} m_\gamma \mathcal{O}^{(s)}_{\gamma\lambda} m_\lambda + \alpha\, r\, c + J(q) - 2 \ll \sigma \; erf\left(\frac{\sigma}{\sqrt{2\alpha r}} \right) \gg_\xi - \sqrt{\frac{\alpha r}{2\pi}} \ll \exp\left(-\frac{\sigma^2}{2\alpha r} \right) \gg_\xi \right) . \tag{2.12}$$

Here $\mathcal{O}^{(s)}_{\gamma\lambda}$ is the matrix element (γ, λ) of $\mathcal{O}^{(s)}$; $erf(x)$ denotes the error function and c is $(1-q)\beta$ in the large β limit. As we have already mentioned, the angular brackets indicate an average over the views. The function $J(q)$ is given by

$$J(q) = -\frac{\alpha}{s} \left[\frac{1 - b + s\, b}{1 - c\,(1 - b + s\, b)} + (s-1) \frac{1 - b}{1 - c\,(1 - b)} \right] \tag{2.13}$$

and σ is defined as

$$\sigma = \sum_{\gamma\lambda}^{s} m_\gamma \mathcal{O}^{(s)}_{\gamma\lambda} \xi^\lambda. \tag{2.14}$$

For given values of the load parameter α and of the strength of the association between a pair of views of the same object b, the values of the overlaps of the stationary state of the network with the representation of the s views of the object are given by the solution of the following equations:

$$m_\nu = \ll \xi_\nu erf\left(\frac{\sum_{\gamma\lambda}^{s} m_\gamma \mathcal{O}^{(s)}_{\gamma\lambda} \xi^\lambda}{\sqrt{2\alpha r}}\right) \gg_\xi \tag{2.15}$$

$$c = \sqrt{\frac{2}{\pi\alpha r}} \ll \exp\left[-\left(\frac{\sum_{\gamma\lambda}^{s} m_\gamma \mathcal{O}^{(s)}_{\gamma\lambda} \xi^\lambda}{\sqrt{2\alpha r}}\right)^2\right] \gg_\xi \tag{2.16}$$

$$r = \frac{1}{s}\left[\frac{1-b+s\,b}{1-c\,(1-b+s\,b)}\right]^2 + \frac{(s-1)}{s}\left[\frac{1-b}{1-c\,(1-b)}\right]^2. \tag{2.17}$$

These equations describe the retrieval properties of one object (let's say object $\mathcal{A}_0$) immersed in a background of another $P_o = \alpha_0 N$ objects (here $\alpha_0 = \frac{\alpha}{s}$ is the load parameter relevant for objects); the effect of these other objects is contained in the parameter r (this is computed trivially from the last equation once c is known) which, as we have said, gives the mean square overlap of all the views of the other objects. The network might also have more complicated states, e.g. a state given by a mixture of two objects. These states are not described by the equations above, but anyway they are not relevant if the cue is sufficiently close to one of the views of object $\mathcal{A}_0$.

3 Results

The simplest way to look for solutions of these equations is to use an iterative procedure, for instance for the s overlaps m_μ we have (the time t denotes the iteration number):

$$m_\mu(t+1) = \ll \xi_\mu \, erf\left(\frac{\sum_{\gamma\lambda}^{s} m_\gamma(t)\mathcal{O}_{\gamma\lambda}^{(s)}\xi^\lambda}{\sqrt{2\alpha r}}\right) \gg_\xi . \tag{3.18}$$

As an initial condition, the overlaps are taken $m_\mu = 0$, except one of them, e.g. m_{μ_0}, that is given a value close to one. The iterations should proceed until a fixed point of this dynamics is reached. However we have noticed that this naive method does not work because it either fails to reach a fixed point or if it does find one, one can not be sure of its stability. For this reason we have preferred to work directly with the equations that result from performing a steepest descent of the free energy. The reader can find a brief description of the method in the Appendix.

The solution of eqs.(2.15-2.17) shows that the object phase is indeed present in a region of the parameter space defined by α and b. As expected it appears when the coupling between views b is strong enough to destabilize the view phase and α is small enough to prevent the noise coming from the storage of an $O(N)$ number of objects from spoiling the retrieval of one of them. For a fixed value of b, one can compute the value of α where the transition from the object to the spin-glass phase takes place. The result is shown in Table 1 for $b = 0.80$ for several values of s, up to eleven views per object. One can see there that the capacity of the network decreases slowly as s increases.

s	$\alpha_0\ s$
3	0.087
5	0.081
7	0.077
9	0.076
11	0.073

Table 1: Capacity (critical value of the load parameter $\alpha = \alpha_0\ s$) of the object phase for several values of the number of views s.

The complete phase diagram of the model is presented in Fig. 5 for a fixed number of views ($s = 5$). The view phase where the views themselves define attractors appears at the left lower corner of the figure; these states disappear at large α when the network makes the usual transition to the spin-glass phase and at large b where the system reaches the object phase. The capacity in terms of the number of objects that can be stored in the network is given by the value of the almost horizontal line in Fig. 5 divided by s.

It is interesting to note that the object phase exists even for b smaller than the values shown in this figure. It is indeed present even for $b = 0$: the standard Hopfield model at low values of $1/\beta$ already has a set of stable symmetric phases where the state of the system has a uniform overlap with an odd number of patterns. In that case however they are not relevant because one is interested in the retrieval of the *patterns* (that we call here the views) and the symmetric states do not show up when the stimulus is close to one of them. As b is increased it will reach a value where one pattern can not support an attractor by itself and then the symmetric (object) states will emerge giving rise to the (object) phase we were looking for.

One consequence of this scenario is that the model can only support objects defined by an odd number of views; however one expects this to be the result of the artificial aspects of the model. In fact, as is discussed in Amit, Gutfreund and Sompolinsky (1985) the properties of these solutions depend on the distribution of the patterns ξ_j^ν . In a more realistic model, with the state of neurons defined in terms of their firing rates we expect this restriction to disappear.

4 Discussion

In this work we have shown that invariant object recognition is feasible in attractor neural networks. The system is able to store and retrieve in a view invariant way an extensive number of objects, each defined by a finite set of views. What is implied by extensive is that the number of objects is proportional to the size of the network. The crucial factor that defines this size is the number of connections per neuron. In the case of the fully connected networks considered in this paper, the size is thus proportional to the number of neurons. To be more specific, the number of objects that can be stored is 0.081 $N/$ 5, when there are five views of each object. The number of objects is 0.073 N / 11, when there are eleven views of each object. This is an interesting result in network terms, in that s views each represented by an independent random set of active neurons, can in the network described, be present on the same "object" attraction basis. It is also an interesting result in neurophysiological terms, in that the number of objects that can be represented in this network scales linearly with the number of recurrent connections per neuron.

Although the explicit numerical calculation was done for a rather small number of views per object (up to 11), the basic result that the network can support this kind of phase is expected to hold for any number of them (the only requirement being that it does not increase with the number of neurons). This is of course enough: once an object is defined by a set of views, when the network is presented with a somewhat different stimulus, or an interpolation of some of the views, or a noisy version of one of them it will still be in the attraction basin of the object attractor.

Some of the assumptions taken to simplify the problem are not relevant in what regards the existence of this regime. This is the case of the assumption that all pairs are coupled to each other with the same strength. Choosing them with a certain distribution will only change the details of where the object phase appears.

Here we have been mainly concerned with the retrieval properties of the network. Because of that we started our analysis from a synaptic matrix where the associations between the views had already been learnt. One can consider however how this matrix is built through a learning procedure that takes place as one object is seen from different perspectives or under different conditions. Learning could proceed as follows: since a fixed view is normally seen for longer times than transitions to another view, it is natural to assume that associations of one view with itself will be built up first while those between different views appear later. This means that the association between views will happen in the presence of attractors for each view. This is possible because at this stage of learning the strength of the association between pairs is still small and, as we have shown, the network will sit in a view phase where the attractors are defined by the views themselves. Under these conditions when the stimulus corresponds to, say, view v_1 the network will reach its attractor and will maintain sustained activity for about 300 ms (Rolls and Tovee, 1994). But within this time scale the object can already present a different view v_2 as the input stimulus. For a short period of time, the neuronal firing may be in a state which reflects both v_1 and v_2. During this time, the appropriate synaptic modification could take place. After these two views are seen consecutively several times, the contribution made by this particular process to the synaptic efficacies will be higher than the threshold required to make a transition to the object phase, and views will start to be recognized as a single object. For a different learning protocol this mechanism has been studied in some detail by Amit and Brunel (1995) and Brunel (1996).

The model emphazises the role of the attractors to implement the association between views but other neurophysiological mechanisms could also participate to reinforce the association. As mentioned before the unbinding of glutamate by the NMDA receptors lasting for about 100ms could produce this effect. Let us also notice that the trace rule used in Foldiak (1991) and Wallis and Rolls (1996) is just one possible way to describe mathematically any of these memory effects.

A study of invariant object representations using recurrent networks has also been done by O'Reilly and Johnson, 1994. Although this is an interesting work, the analysis is mainly numerical and the capacity properties of the model are not fully understood. Invariant representation of faces in the context of attractor neural networks has been discussed by Bartlett and Sejnowski (1996) in terms of a model where different views of faces are presented in a fixed sequence (Griniasty, Tsodyks and Amit, 1993). This is not however the general situation; normally any pair of views can be seen consecutively and they will become associated. The most general version of the model presented in this work contemplates this possibility. [3] The synaptic matrix in eq.(2.3) refers to this situation, although later on we have found it more convenient to study the simpler case where all the views are coupled to each other with the same strength.

We wish to note the different nature of the invariant object recognition problem studied here, and the paired associate learning task studied in (Miyashita and Chang, 1988; Miyashita,

[3] After this work was submitted, we knew that in a more recent publication Bartlett and Sejnowski (1997) have relaxed the condition on the presentation of the views in a fixed sequence. In this respect their recurrent synaptic matrix is closer to ours.

1988; Sakai and Miyashita, 1991). In the invariant object recognition case no particular learning protocol is required to produce an activity of the IT cells responsible for invariant object recognition maintained for 300 ms. The learning can occur rapidly, and the learning occurs between stimuli (e.g. different views) which occur with no intervening delay. In the paired associate task, the monkeys must learn to associate together two stimuli that are separated in time (by a number of seconds), and this type of learning can take weeks to train. During the delay period the sustained activity is rather low in the experiments, and thus the representation of the first stimulus that remains is weak, and can only poorly be associated with the second stimulus. However, formally the learning mechanism could be treated in the same way as we have used here for invariant object recognition. The experimental difference is just that in the paired associate task used by Miyashita et al, it is the weak memory of the first stimulus that is associated with the second stimulus. In contrast, in the invariance learning, it would be the firing activity being produced by the first stimulus (not the weak memory of the first stimulus), that can be associated together.

The mechanisms described here would apply most naturally when a small number of representations need to be associated together to represent an object. One example is associating together what is seen when an object is seen from different perspectives. Another example is scale, with respect to which neurons early in the visual system tolerate scale changes of approximately 1.5 octaves, so that the whole scale range could be covered by associating together a limited number of such representations (see Rolls, 1994, 1996). The mechanism would not be so suitable when a large number of different instances would need to be associated together to form an invariant representation of objects, as might be needed for translation invariance.

For the latter, we propose a solution in a multilayer network, with a local solution being implemented at each stage (Rolls, 1994, 1995; Wallis and Rolls, 1996). We have envisaged the local solution for translation invariance at each stage as being performed by a trace rule implemented between the inputs to a stage and the postsynaptic neurons in a stage. However, both types of mechanism, implemented that is in the feedforward connections, or in the recurrent collateral connections, could contribute (separately or together) to achieving invariant representations. Part of the interest of the approach described in this paper is that it allows analytic investigation, and this is what has been introduced in this paper.

Acknowledgements

This research was supported by the Medical Research Council, PG8513790; by an E.U. Human Capital and Mobility grant CHRX-CT92-0063 and by a Spanish grant PB96-47.

Appendix

Here we explain the way eqs.(2.15-2.17) were solved. A straightforwrd way to perform a steepest descent on the free energy given in eq.(2.12) is to update the order parameters according to the variation of the free energy with respect to m_μ, r and c. The complication of this approach is that it yields $P+2$ coupled equations, even when we know that r and c are trivially related at the fixed point, as eq(2.17) shows. One can make convergence faster by imposing this relation as a constraint on the dynamics by means of a Lagrange multiplier. The algebra is rather simple and one finally finds the following equations for the updating of m_μ, r and c:

$$\delta m_\nu = -\eta(1-a)\left[m_\nu - \ll \xi^\nu erf\left(\frac{\sigma}{\sqrt{2\alpha r}}\right) \gg_\xi\right] + a\left[M - \ll Y_\xi \, erf\left(\frac{\sigma}{\sqrt{2\alpha r}}\right) \gg_\xi\right] \quad (4.19)$$

$$\delta r = -\eta\frac{\alpha}{2}\left[c - \sqrt{\frac{2}{\pi\alpha r}} \ll \exp\left(-\frac{\sigma^2}{2\alpha r}\right) \gg_\xi\right] + \eta\,\lambda \quad (4.20)$$

$$\delta c = -\eta\frac{\alpha}{2}\left[r - K(c)\right] - \eta\,\lambda\frac{\delta K(c)}{\delta c}. \quad (4.21)$$

In these equations we have defined:

$$M = \sum_{\lambda=1}^{s} m_\lambda \quad (4.22)$$

$$Y_\xi = \sum_{\lambda=1}^{s} \xi^\lambda \quad (4.23)$$

$$K(c) = \frac{1}{s}\left[\frac{1-a+s\,a}{1-c(1-a+s\,a)}\right]^2 + \frac{(s-1)}{s}\left[\frac{1-a}{1-c(1-a)}\right]^2, \tag{4.24}$$

and λ is chosen in such a way that if $r = K(c)$ holds at a given time (in particular at $t = 0$) then it also holds at the next time step:

$$r + \delta r = K(c + \delta c). \tag{4.25}$$

Although it is correct to solve eqs (4.19-4.21) to see if the invariant recognition phase is reached upon presentation of one view, the connection between these equations and the dynamics defined by eq (2.8) is subtle. Eq 2.8 refers to the evolution of a microscopic state. It decreases systematically the value of the Hamiltonian, eq (2.6). On the other hand, eqs (4.19-4.21) decrease the free energy. Strictly speaking the fixed point is not the same. This is because once a number $O(N)$ of views have been stored in the synaptic matrix, there will be many metastable states with a large (i.e. close to one) overlap with a given stored view. Eqs (4.19-4.21) reach a fixed point that describes average properties of these states highly correlated to the stored views. Eq (2.8) reaches one of these states. If the initial condition is modified (but still starting from a state close to one of the views), the final state will probably be a different one, although it will still have a large overlap with the stored pattern. A study of the number of metastable states in an associative memory network has been done by E. Gardner (1986).

Bibliography

Amit, D., Gutfreund, H. and Sompolinsky, H. (1985) Spin-glass models of neural networks. *Phys. Rev.* **A 32**, 1007-1018.

Amit, D. (1988) *Modeling brain function.* Cambridge Univ. Press.

Amit, D. and Brunel, N. (1995) Learning internal representations in an attractor neural network with analogue neurons. *NETWORK* **6**, 359-388.

Bartlett, M. S. and Sejnowski, T.J. (1996) Learning viewpoint invariant representations of faces in an attractor network. Communication presented at the 18th Cognitive Science Meeting, San Diego, CA.

Bartlett, M. S. and Sejnowski, T.J. (1997) Viewpoint invariant face recognition using independent component analysis and attractor networks. In Mozer M., Jordan M. and Petsche T. (Eds.), *Advances in Neural Information Processing Systems* **9**, MIT Press, Cambridge, MA.

Brunel, N. (1996) Hebbian learning of context in recurrent neural networks. *Neural Computation* **8**, 1677.

Cugliandolo, L. F. and Tsodyks, M. V. (1994) Capacity of networks with correlated attractors. *Journal of Physics* **A 27**, 741-755.

Foldiak, P. (1991) Learning invariance from transformation sequences. *Neural Computation* **3**, 193-199.

Gardner, E. (1986) Structure of metastable states in the Hopfield model. *Journal Physics* **A 19**, L1047-L1052.

Gardner, E. (1988) The space of interactions in neural network models. *Journal Physics* **A 21**, 257-270.

Griniasty, M., Tsodyks, M. V. and Amit, D. J. (1993) Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation* **5**, 1.

Gross, C. G., Desimone, R., Albright, T. D. and Schwartz, E. L. (1985) Inferior temporal cortex and pattern recognition. *Exp. Brain Res. Suppl.* **11**, 179-201.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Natl. Acad. Sci. USA* **79**, 2554-2558.

Mezard M., Parisi G. and Virasoro M.A. (1987) *Spin Glass Theory and Beyond.* World Scientific Lecture Notes in Physics Vol. 9.

Miyashita Y. (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* **335**, 817-820.

Miyashita, Y. and Chang, H. S. (1988) Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* **331**, 68-70.

O'Reilly, R.C. and Johnson, M.H. (1994) Object recognition and sensitive periods: a computational analysis of visual imprinting. *Neural Computation* **6**, 357-389.

Rolls,E.T. (1984) Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiology* **3**, 209-222.

Rolls, E. T. (1992) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Phil. Trans. Roy. Soc.* **335**, 11-21.

Rolls, E. T. (1994) Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes* **33**, 113-138.

Rolls, E. T. (1995) Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Res.* **66**, 177-185.

Rolls, E. T. (1996a) Roles of Long Term Potentiation and Long Term Depression in neuronal network operations in the brain. In: *Cortical Plasticity: LTP and LTD.* Eds. M. S. Fazeli and G. L. Collingridge. Bios: Oxford.

Rolls, E. T. (1996b) A neurophysiological and computational approach to the functions of the temporal lobe cortical visual areas in invariant object recognition. In: *Computational and Biological Mechanisms of Visual Coding*, eds. L. Harris and M. Jenkin. Cambridge University Press: Cambridge.

Rolls, E. T., Booth, M. C. A. and Treves, A. (1996) View-invariant representations of objects in the inferior temporal visual cortex. *Society for Neuroscience Abstracts* **22**

Rolls, E. T., Tovee M. J. (1994) Processing speed in the cerebral cortex, and the neurophysiology of visual backward masking. *Proc. Roy. Soc.* **B 257**, 9-15.

Rolls, E. T., Tovee M. J., Purcell D. G., Stewart AL, Azzopardi P (1994) The responses of neurons in the temporal cortex of primates, and face identification and detection. *Exp. Brain Res.* **101**, 474-484.

Rolls, E.T. and Treves, A. (1997) *Neuronal Networks and Brain Function.* Oxford University Press.

Sakai, K. and Miyashita Y. (1991) Neural organisation for the long-term memory of paired associates. *Nature* **354**, 152.

Sutton, R. S. and Barto, A. G. (1981) Towards a modern theory of adaptive networks: Expectation and prediction. *Psychological Review* **88**, 135-170.

Tanaka K, Saito C, Fukada Y, Moriya M (1990) Integration of form, texture, and color information in the inferotemporal cortex of the macaque. In: *Vision, Memory and the Temporal Lobe*, ch. 10, pp. 101-109. Eds. E. Iwai and M. Mishkin. Elsevier: New York.

Tsodyks, M. V. and Feigel'man, M. V. (1988) The enhanced storage capacity in neural networks with low activity level. *Europhysics Lett.* 101-105.

Wallis, G. Rolls, E. T. and Foldiak, P. (1993) Learning invariant responses to the natural transformations of objects. In: *International Joint Conference on Neural Networks*, pp. 1087-1090, vol. 2.

Wallis,G. and Rolls,E.T. (1997) Invariant face and object recognition in the visual system. *Progress in Neurobiology* **51**, 167-194.

FIGURE CAPTIONS

Figure 1. A trace learning rule is implemented in the feedforward inputs to a non-recurrent network.

Figure 2. A trace or pairwise associative learning rule is implemented in the recurrent collateral synapses of an autoassociative memory.

Figure 3. A schematic illustration of the first type of associations contributing to the synaptic matrix considered (see text).

Figure 4. A schematic illustration of the second and main type of synaptic matrix considered (see text).

Figure 5. The phase diagram of the model for $s = 5$. Region I is the standard view phase where the views themselves are retrieved correctly. In this region objects are not stored in a view invariant fashion. Region II is the phase where objects are stored as attractors, any stimulus close to some of the views will elicit the same response. Finally in region III, the spin-glass phase, no storage of information is possible.

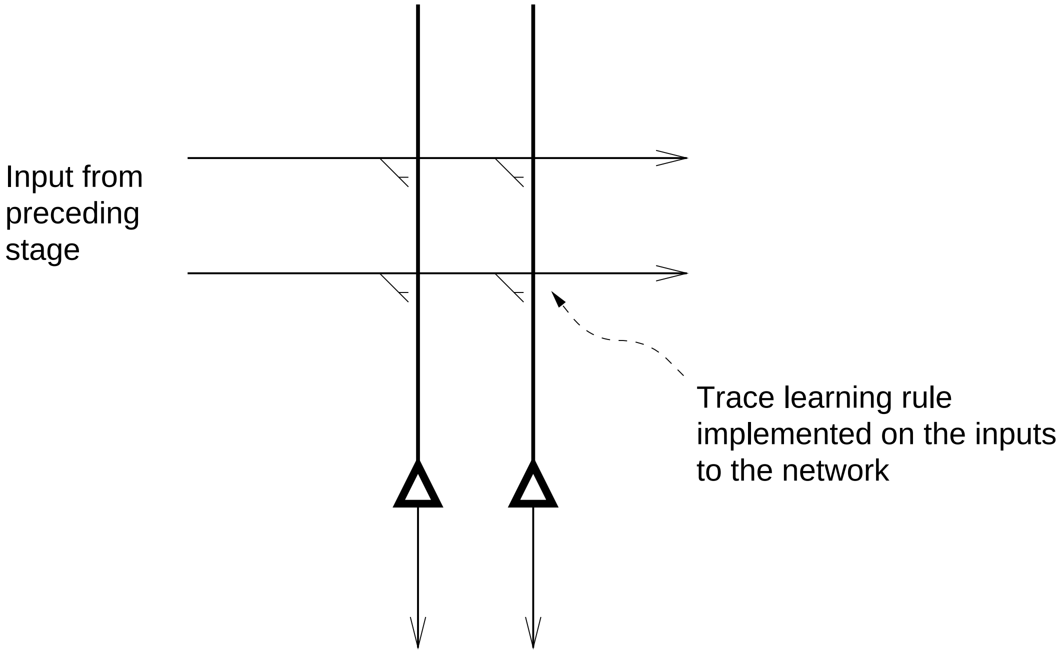
Input from
preceding
stage
Trace learning rule
implemented on the inputs
to the network

Figure 1:

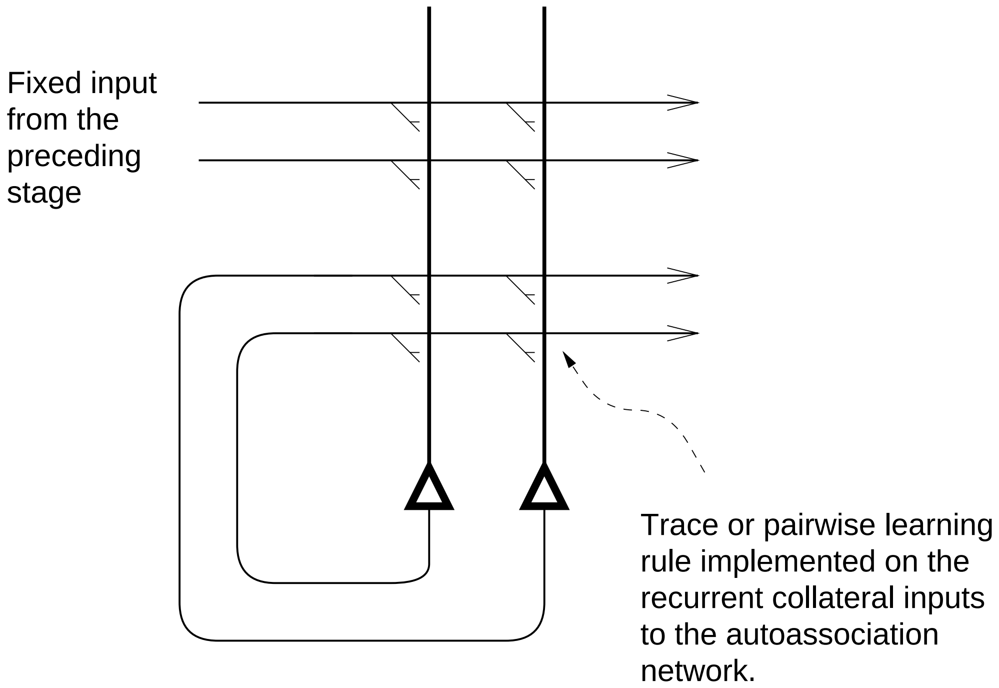
Fixed input
from the
preceding
stage
Trace or pairwise learning
rule implemented on the
recurrent collateral inputs
to the autoassociation
network.

Figure 2:

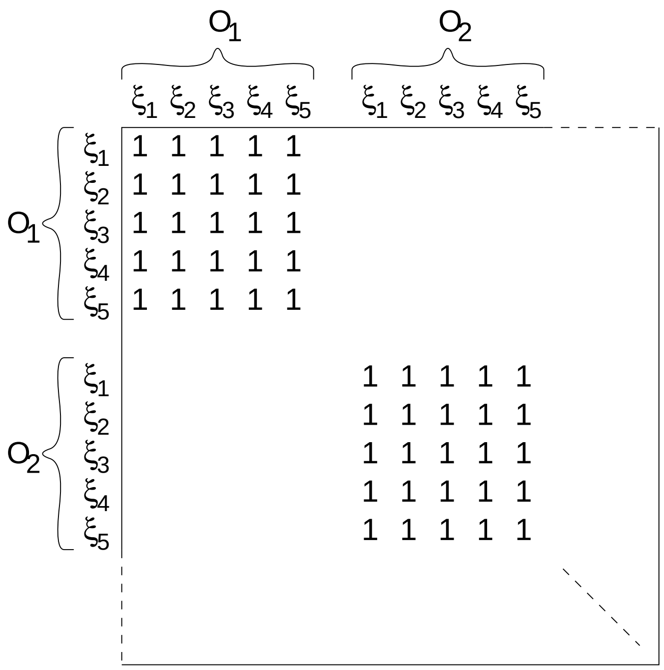

Figure 3:

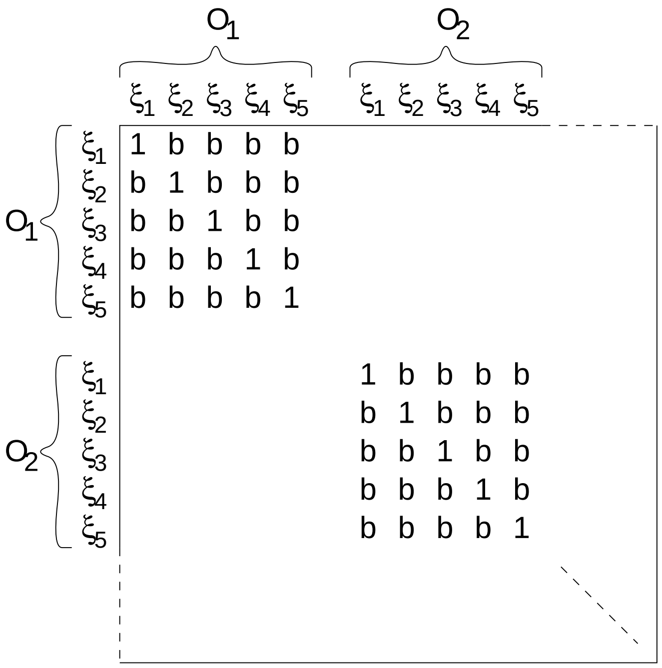
O1
O2
ξ1 ξ2 ξ3 ξ4 ξ5
ξ1 ξ2 ξ3 ξ4 ξ5
O1
ξ1
ξ2
ξ3
ξ4
ξ5
1 b b b b
b 1 b b b
b b 1 b b
b b b 1 b
b b b b 1
O2
ξ1
ξ2
ξ3
ξ4
ξ5
1 b b b b
b 1 b b b
b b 1 b b
b b b 1 b
b b b b 1

Figure 4:

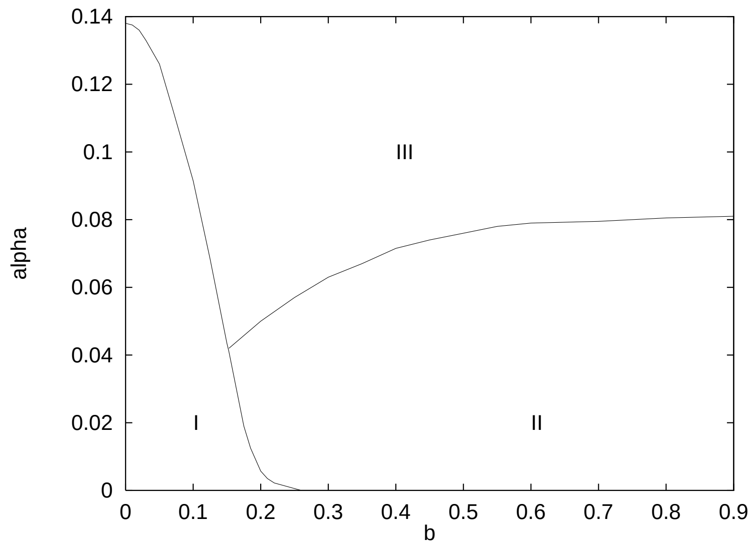
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
alpha
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
b
I
II
III

Figure 5: