A Recurrent Model of Transformation Invariance by Association

Martin C.M. Elliffe [a], Edmund T. Rolls [a,1,2], Néstor Parga [b,2,3], and Alfonso Renart [b,3]

[a] *Department of Experimental Psychology Oxford University, Oxford OX1 3UD, United Kingdom*

[b] *Departamento de Física Teórica, C-XI Universidad Autonoma de Madrid, 28049 Madrid, Spain*

Reprint requests should be sent to E.T. Rolls, Dept. of Experimental Psychology, South Parks Rd, Oxford OX1 3UD, United Kingdom.

Abstract

This paper describes an investigation of a recurrent artificial neural network which uses association to build transform-invariant representations. The simulation implements the analytic model of Parga and Rolls [16] which defines multiple (e.g. "view") patterns to be within the basin of attraction of a shared (e.g. "object") representation.

First, it was shown that the network could store and correctly retrieve an "object" representation from any one of the views which define that object, with capacity as predicted analytically.

Second, new results extended the analysis by showing that correct object retrieval could occur where retrieval cues were distorted; where there was some association between the views of different objects; and where connectivity was diluted, even when this dilution was asymmetric. The simulations also extended the analysis by showing that the system could work well with sparse patterns; and showing how pattern sparseness interacts with the number of views of each object (as a result of the statistical properties of the pattern coding) to give predictable object retrieval performance. The results thus usefully extend a recurrent model of invariant pattern recognition.

Key words: Object Recognition; Invariance; Recurrent Networks; Attractor Networks; Associative Learning; Sparse Coding; Invariant Visual Representations.

---

1 Introduction

Single neurons with responses which are relatively invariant with respect to the position; size; view (rotation in depth); and other transformations of an object or face are present in the primate temporal visual cortical areas (see e.g. [11,27,17–20,5]). How could such invariant representations be formed? One suggestion is based on the temporal statistics with which objects are normally experienced in the visual world: the identity of the presented object changes relatively slowly by comparison with the faster-changing transformation (e.g. view) of that object. Thus, if an on-line learning rule could include a short-term memory trace, such that views presented close to each other in time (e.g. at successive training steps) could be associated together to gain similar representations, so transform-invariant representations could develop ([8,17,10,20]). Such representations would be unlikely to span multiple different objects, given that views of different objects are unlikely to repeatedly occur close together in time.

Földiák ([8]) showed translation invariance over a one-dimensional input array to be possible for a simple competitive network with an associative Hebb learning rule including a decaying trace of previous neuronal activity. Rolls ([17,18,22]) suggested that invariant representations could be formed in the visual system for two-dimensional images using a similar learning rule in a multi-layer feed-forward architecture, incorporating many neurobiological properties such as limited inter-layer connectivity and soft competition. This system was simulated by Wallis and Rolls ([34]) and was shown to be capable of performing translation- and view-invariant object recognition.

In an approach to the learning of associations between adjacent pairs of patterns in a sequence, Griniasty et al ([10]) introduced a recurrent model (a single-layer auto-associator, or "attractor" network) with a fixed stimulus presentation order and a learning rule which incorporated correlation between activity and the immediate preceding stimulus (i.e. a form of trace), and showed that the resultant attractors included partial correlation with temporally-proximal stimuli. Parga and Rolls ([16]) generalised this approach to include equal association between *all* different views of the same object, and defined the resultant model in analytically-tractable form. Their analysis showed that the strength of association between the views of each object was a critical parameter: at low association strength the model exhibited "view phase" behaviour, such that presentation of a (possibly distorted) view would result in retrieval of that view; while at higher strengths the model exhibited

[1] MRC Program Grant PG8513790.
[2] McDonnell-Pew Center for Cognitive Neuroscience, Oxford University: European Network Grant.
[3] Spanish Ministry of Education and Culture Grant PB96–47.

"object phase" behaviour, such that presentation of any view of some object would result in retrieval of a shared representation of that *object*.

A particular advantage of the Parga and Rolls model is the analytic investigation it allows, and the quantitative parameter predictions they thus provide. For these reasons, we explore it further in this paper. In particular, we not only confirm the results predicted by the analysis, but also extend the analytic results in a number of ways, several of which are relevant to how such a system might operate in the brain.

There have been many previous contributions addressing the problem of networks capable of achieving recognition with respect to various transformations in the input space. All of them look for a single representation reached by the network when a wide class of stimuli, to be identified with local views of an object, are presented as input. For example, Dotsenko [7] proposed a network in which rotated, translated or rescaled patterns presented to the network, could be recognized making use of the neuronal thresholds as additional dynamical variables. In this way, patterns very different from the stored ones as measured by Hamming distance, could be retrieved and represented by a persistent configuration of activity if they were related to any of the stored patterns through the type of transformations listed above. From a different perspective, von der Malsburg and Bienenstock ([33]) also proposed a means of storing invariant representations by changing the stored entities from patterns of activity to graphs containing information about the correlation in the activity of the neurons. The storage and retrieval of such graphs automatically implements invariant recognition of patterns by the identification of graphs which are isomorphic. Our approach is more general, since it assumes no *a priori* relation between the different views of a given image, which are, in fact, taken as independent. This independence between the patterns associated with each of the local views is, however, not essential for the model to work and is only assumed for simplicity. If the local views corresponding to a given image were correlated, the associations in the synaptic efficacies connecting the neurons active in each of them would be easier to create, although probably the storage capacity of the network would decrease. The model demonstrates that the mechanism transforming temporal into spatial correlations provides invariant representations even in the *worst case scenario* in which views corresponding to images close in time are represented by independent patterns.

It should be stressed that the type of generalization that our model accomplishes by creating a single attractor that can be reached from any of the local views of an object is very different from the generalization over short Hamming distances achieved by conventional associative memory systems. In the model, although the object attractor and the attractors corresponding to the local views are in fact *close* in Hamming distance, this fact is immaterial. What is relevant is that the patterns corresponding to the local views can be as different as one wants. The proposed learning mechanism creates a single representation for the whole object which is based on the perceived partial representations and their temporal correlations as experienced by the animal. Finally, it should also be noted that the model described here not only achieves invariant recognition as it is meant to, but it also employs a plausible learning procedure based on mechanisms that have the support of experimental evidence ([14,35]).

The remainder of this paper investigates the performance of the model by empirical means. Sections 2 and 3 of this paper provide the theoretical basis for the model and its empirical performance measurement respectively. Section 4 compares empirical model performance with that calculated analytically, while sections 5 through 8 present empirical results showing the model's tolerance to variation of, respectively, cue/view distortion; association strength; connectivity dilution and the symmetry of that dilution; and pattern sparseness. A discussion interprets the results presented, with appendices detailing the application of the model to alternative binary pattern domains, and presenting a signal-to-noise analysis consistent with the empirical sparseness results of section 8.

2 Model Definition

The model described by Parga and Rolls [16] provides transform-invariant "object" fixed-point attractors in a Hopfield-like network [12]. Each such object is represented by s "views", each of which is a pattern of N binary components (taking values 0 and 1, for notational convenience, but see Appendix A).

Following many previous researchers (e.g. [1,31]), "neurons" in the model are dynamical binary variables, with values $\{\mathcal{V}_j(t)\}, j = 1, ..., N$ at time t given by :-

$$\mathcal{V}_j(t+1) = \Phi(\sum_{j \neq i} \mathcal{J}_{ij} \mathcal{V}_j(t) + h_i(ext)) \tag{1}$$

where Φ is a binary transduction function (equation 2); $\mathcal{J}_{ij}$ is the connection strength between neurons i and j; and $h_i(ext)$ is any external input to neuron i. Note that the implementation uses immediate decay of the external input: it is used for a single timestep, and then set to zero (i.e. no input clamping).

Transduction function Φ is simply the sign function :-

$$\Phi(I) = \begin{cases} -1 & I < \theta \\ +1 & I \geq \theta \end{cases} \tag{2}$$

where threshold $\theta = 0$.

Views consist of N random binary variables, such that :-

$$P(\eta_i^{\beta\mu}) = a\delta(\eta_i^{\beta\mu} - 1) + (1-a)\delta(\eta_i^{\beta\mu}) \tag{3}$$

$\forall \beta,\ \mu,\ i$ where $i = 1, ..., N$ indexes the neuron; $\mu = 1, ..., s$ indexes the view within an object (as does ν); $\beta = 1, ..., P_o$ indexes the object (as does γ); and a is the pattern sparseness (for [0,1] binary patterns, simply the proportion of components taking value 1, but see [24]), which corresponds to the symbol w used by Parga and Rolls. The implementation ensures that each of the total number of views ($L \equiv sP_o$) consists of exactly the same number of 1s ($N_a' = \lfloor aN + 0.5 \rfloor$, the rounded equivalent of sparseness a applied to all N cells). The actual procedure used set the first N_a' components of each pattern to 1, the remaining $N - N_a'$ components to 0, and then randomly sorted those components.

The invariant recognition model of Parga and Rolls is unique, however, in how it specifies the synaptic strengths. This implements the strength of association between the different views of an object, with detail :-

$$\mathcal{J}_{ij} = \frac{d_{ij}}{a(1-a)Nd} \sum_{\beta=1}^{P_o} \sum_{\mu,\nu=1}^{S} (\eta_i^{\beta\mu} - a)\chi^{\mu\nu}(\eta_j^{\beta\nu} - a) \tag{4}$$

given $\chi^{\mu\nu}$ as the strength of the association between views (μ and ν) of the same object (β) :-

$$\chi^{\mu\nu} = \delta^{\mu\nu} + b(1^{\mu\nu} - \delta^{\mu\nu}) \tag{5}$$

$$\hat{\chi} = \left\{ \underbrace{\begin{matrix} 1, & b, & ..., & b, & b \\ b, & 1, & ..., & ..., & b \\ ..., & ..., & 1, & ..., & ..., \\ b, & ..., & ..., & 1, & b \\ b, & b, & ..., & b, & 1 \end{matrix}}_{s} \right\}$$

The implementation described here in fact extends the generality of this scheme by allowing specification of each such association strength, termed :-

- b_v between each view and itself (1, in all simulations here);
- b_o (or simply b, and used interchangeably in this paper) between each view and all other views of the same object; and
- b_d between each view and all other views of different objects.

Connectivity probability d (termed "dilution", though it is in fact dilution's complement, with full connectivity at $d = 1$) is included by random variable d_{ij} distributed according to :-

$$P(d_{ij}) = d\delta(d_{ij} - 1) + (1 - d)\delta(d_{ij}) \tag{6}$$

Connectivity is full ($d = 1$) unless stated otherwise.

Network loading is defined by :-

$$\alpha \equiv \frac{sP_o}{Nd} \tag{7}$$

The remaining unspecified parameter is the external cue, $h_i(ext)$, which is a distorted version of one of the views, albeit retaining the constant number of 1s (N'_a, as above). Distortion was applied by first calculating the number of 1s which must be common between the view and its distortion ($N_1 = \lfloor 0.5 + Na^2 + r(Na - Na^2)\rfloor$) for some desired cue/view Pearson product-moment correlation r, and thus the number of pattern components which must swap values ($N_\Delta = N'_a - N_1$). A distorted view is thus a direct copy of the original except with N_Δ randomly selected original 0s set to 1 and N_Δ randomly selected original 1s set to 0. The actual distorted external cue used is now $h_i^{\beta\mu}(ext) = g(\tilde{\eta}_i^{\beta\mu})$ where $g()$ is a function which translates values from the pattern domain to those appropriate for the transduction function Φ, and $\tilde{\eta}^{\beta\mu}$ is view μ of object β distorted as detailed above. Given Φ as the sign function (equation 2), the translation function is :-

$$g(I) = \begin{cases} -1 & I = 0 \\ +1 & I = 1 \end{cases} \tag{8}$$

Note that except where specified otherwise, all cues used were distortion-free, being simply the views themselves subject to translation by $g()$.

Dynamics of the implementation are fully parallel (all neurons calculate their transduction functions simultaneously) and are thus deterministic. Performance was similar when using stochastic dynamics (where the neuron to be updated is chosen at random, optionally with or without replacement during each epoch), but was somewhat less predictable: deterministic dynamics removed a source of inter-trial variability considered unnecessary in this study.

3 Performance Evaluation

The model's theoretical basis shows that the network loading (α) and, particularly, intra-object view association strength (b_o, generically termed b) prescribe the *phase* in which the system operates. Three phases exist :-

I. view: where an attractor exists for each view;
II. object: where an attractor exists for each object, overlapping equally with each view; and
III. spin-glass: where the attractor reached is not correlated with any view of any object.

The criteria for inclusion in each phase are quantified by the use of two metrics, each based on the Pearson product-moment correlation :-

$$r(\mathcal{V}, \mu^{\beta\nu}) = \frac{(\sum_j^N \mathcal{V}_j \eta_j^{\beta\nu}) - N\mu_{\mathcal{V}}\mu_{\eta^{\beta\nu}})}{(N-1)\sigma_{\mathcal{V}}\sigma_{\eta^{\beta\nu}}} \tag{9}$$

where $\mathcal{V}$ is a network state (with mean $\mu_{\mathcal{V}}$ and standard deviation $\sigma_{\mathcal{V}}$) and $\eta^{\beta\nu}$ is the ν^{th} view of object β.

The "View Metric" identifies the view phase using :-

$$M_v = r(\mathcal{V}, \mu^{\beta\nu}) - max(\forall_{\gamma,\mu\neq\beta,\nu} : r(\mathcal{V}, \mu^{\gamma\mu})) \tag{10}$$

which is simply the correlation between the attractor state and the cued object view ($\beta\mu$) less the maximum correlation between the attractor state and any other view. Thus, positive values occur only where the state is more correlated with the cued *view* than with any other view, including other views of the same object.

The "Object Metric" identifies the object phase using :-

$$M_o = min(\forall_\nu : r(\mathcal{V}, \mu^{\beta\nu})) - max(\forall_{\gamma\neq\beta,\nu} : r(\mathcal{V}, \mu^{\gamma\nu})) \tag{11}$$

which is simply the minimum correlation with any view of the cued object (β) less the maximum correlation with any view of any other object. A positive

M_o value thus indicates that the network state is more correlated with *all* views of the cued object than with *any* view of any other object.

The M_v and M_o values reported for entire networks are the mean such values over all cues (one for every view of every object), unless otherwise stated.

The spin-glass phase is generally identified by default: that is, it is assumed to apply where neither of the other metrics positively identifies the phase (i.e. both M_v and M_o are zero).

Note that stability is evaluated by calculating the correlation between the current and previous network states. Where the correlation differs by less than a specified threshold (0.001) for a specified number of iterations (10), the network is deemed to have achieved stability. The number of iterations required to reach a view- or object-retrieval attractor is small (typically ≤ 6 throughout this study), whereas the number of iterations required to reach a spin-glass attractor may be much greater [4]: a pragmatic limit (100) to the number of iterations was therefore used.

4 Analytic/Empirical Performance Comparison

The first empirical investigation was comparison with the analytic results reported by [16]. Simulations were therefore run ranging over a similar $[\alpha, b]$ parameter space, with 5 views for each object (s).

Figure 1 shows the resultant "View" (top-left) and "Object" (top-right) metrics, including a plane with View Metric 0 to highlight positive values. The resultant phase diagram (bottom-left, where I is the view phase; II the object phase; and III the spin-glass phase) is also shown, with the analytic results of Parga and Rolls shown in the bottom-right for comparison. The empirical phase diagram shows boundaries in two forms: solid lines are contours where the metrics reach zero (indicating no retrieval, the analytic phase boundary), while dotted lines are contours where the metrics reach 0.75 of their maximum, provided to indicate the landscape of the space. The points where the space was sampled in the empirical investigations are shown as "."-marks, and are reproduced on the analytical diagram for ease of comparison.

The empirical and analytic results are very similar, with special attention being drawn to the following points.

[4] The exact nature of all high-iteration states is not investigated here: given asymmetrically diluted connectivity, for example, cyclical or chaotic state sequences might be followed. In either case, however, the states are not retrieval attractors, which are the object of this study.

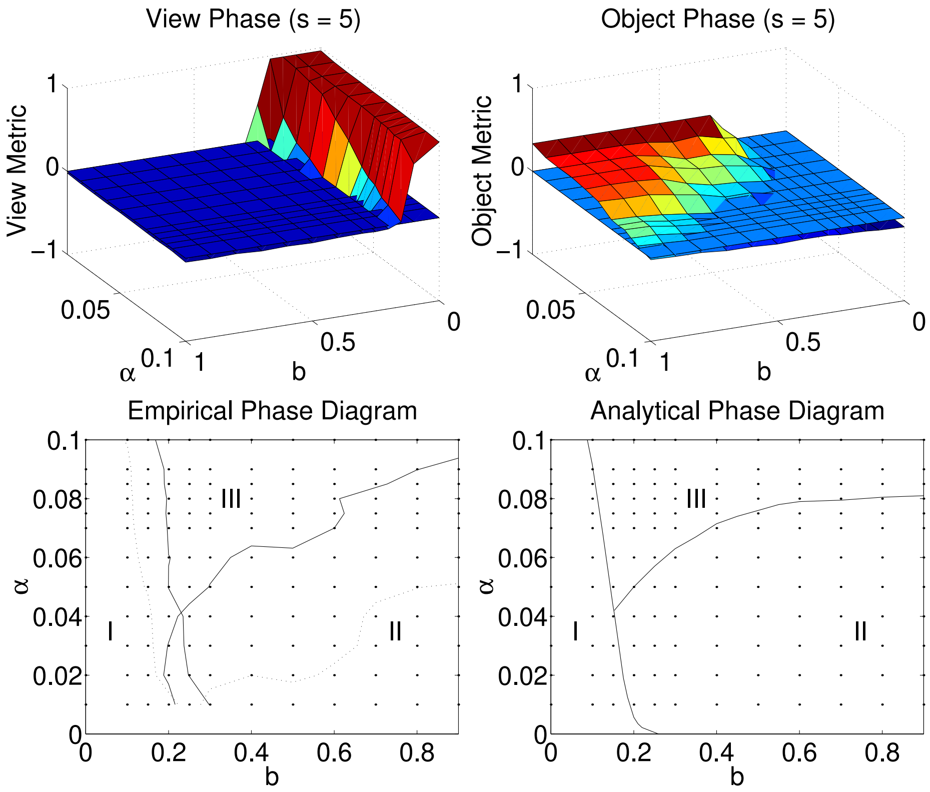

Fig. 1. Model performance given $N = 1000$; $d = 1$; $a = 0.5$; and $s = 5$-view objects ranging across loading (α) and intra-object association strength (b). The individual "View" (top-left) and "Object" (top-right) metrics are shown, together with the resultant empirical phase diagram (bottom-left). For comparison, the analytical phase diagram is also shown (bottom-right).

First, the empirical view phase extends to higher values of b than the analytic equivalent, and indeed the view and object phases appear to overlap. By examination of the "View Metric" results (top-left), however, it is clear that the metric decreases sharply as soon as b exceeds some crucial value: the dotted phase boundary shows the contour where the metric surface passes through 0.75 of its maximum, and is now very close to the analytic boundary. The empirical results show that while the view phase does end abruptly, there is a thin region of parameter space where stable attractor states exist for only some rather than all views. Moreover, the region of apparent phase overlap shows that the empirical environment allows at least one attractor to be shared by several views of the same object, while other views have individual attractors.

Second, the empirical object phase boundary occurs at smaller values of α than the analytic equivalent, though at high b the empirical object phase extends to marginally higher α.

Three issues are relevant to both points :-

1. interpolation: the parameter space has only been sampled at the "."-

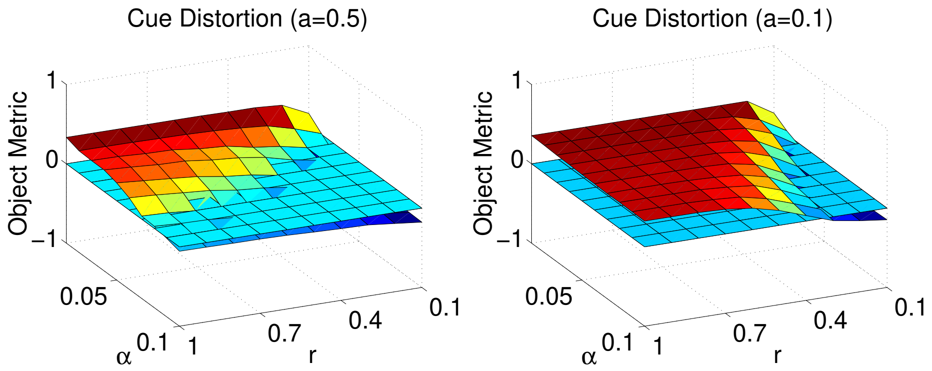

Fig. 2. Cue Distortion. Object metric performance given $N = 1000$; $d = 1$; $b = 0.7$; and $s = 5$ for $a = 0.5$ (left) and $a = 0.1$ (right) across ranges of network loading (α) and cue/view correlation (r).

marks shown. The contours shown are the result of linear interpolation between these sample points, and thus are indicative rather than precise;

2. finite size: the networks used consist of a finite number of neurons ($N = 1000$), and thus some variation from analytic performance is to be expected (i.e. while the pattern generation process is genuinely random, it may well be the case that some views, whether they be within the same or different objects, will be more correlated with each other than with any other view). Note that the position and general form of the phase transitions has been confirmed with larger networks ($N = 3000$), but that production of a fully-sampled phase diagram, for example, is prohibitively computationally expensive. The value of N at which finite size effects are effectively eliminated for this paradigm remains unknown; and
3. treatment of non-retrieval states: the metrics both return 0 (zero) where a cue presentation fails to result in a retrieval state, since there is no evidence for inclusion in either the view or object phase. The results shown are the mean metric values over all presented cues, however, and thus the empirical phase boundaries show the outer limits of the parameter space where evidence to support each phase has been found.

The dotted line at 0.75 of the maximal object metric has been included purely for consistency with the "View Phase" presentation.

5 Cue Distortion

The cues presented for retrieval may vary, being imperfect (i.e. distorted) versions of object views. The network was shown to be tolerant to these variations, with the following quantitative results.

Figure 2 shows the effect of different quantities of distortion of the view-based cues, expressed in terms of the correlation between view and cue, across a range of network loading (α). Eight cues were generated for each view at each required correlation level, with the values shown being the mean of the resultant object metrics across all such presentations. Constant parameter values include the number of views per object ($s = 5$); the intra-object association strength ($b = 0.7$, an arbitrary choice); and the level of pattern sparseness ($a = 0.5$ for the left-hand graph; $a = 0.1$ for the right).

The results show that as either network loading increases or cue/view correlation decreases, so the object metric values generally decrease. However, the right-hand graph (low sparseness) shows particularly clearly that at any constant network loading, the cue need only be above some critical correlation threshold for performance to be maximal: as network loading increases, so this correlation threshold also increases. The important result shown is thus that cues need not be perfect versions of the object views, with the model able to generalise across novel cues provided that those cues are correlated sufficiently closely for the given network loading.

Note that the effect of varying pattern sparseness (a) is investigated systematically in section 8.

6 Association Strength

Figure 3 shows both view (left) and object (right) performance across ranges of the association strength parameters b_o (between different views of the same object) and b_d (between views of different objects) for a variety of network sizes and loadings. The top row shows results for $N = 1000$ and $\alpha = 0.025$; the middle row for $N = 2000$ and $\alpha = 0.025$; and the bottom row for $N = 1000$ and $\alpha = 0.05$. Note that strength b_v (between each view and itself) is held constant at 1.

Both the view (left) and object (right) graphs show that small inter-object association strength (b_d) is required to gain positive metric values and, as shown in section 4, increasing intra-object association strength (b_o, or generic b) moves the model from the view to the object phase.

Inter-object association strength b_d introduces a term in the mean current of each neuron which is of order $b_d\sqrt{N}$, and will tend to destabilise the attractors. Comparison of the top row graphs ($N = 1000; \alpha = 0.025$) with either the middle ($N = 2000; \alpha = 0.025$) or bottom ($N = 1000; \alpha = 0.05$) row shows that increasing network size (N) or network loading (α) decreases the size of both phases in $[b_o, b_d]$-space. In the analytically-tractable limit of infinite N, no stable attractors exist where $b_d \neq 0$: the results here are thus an effect of

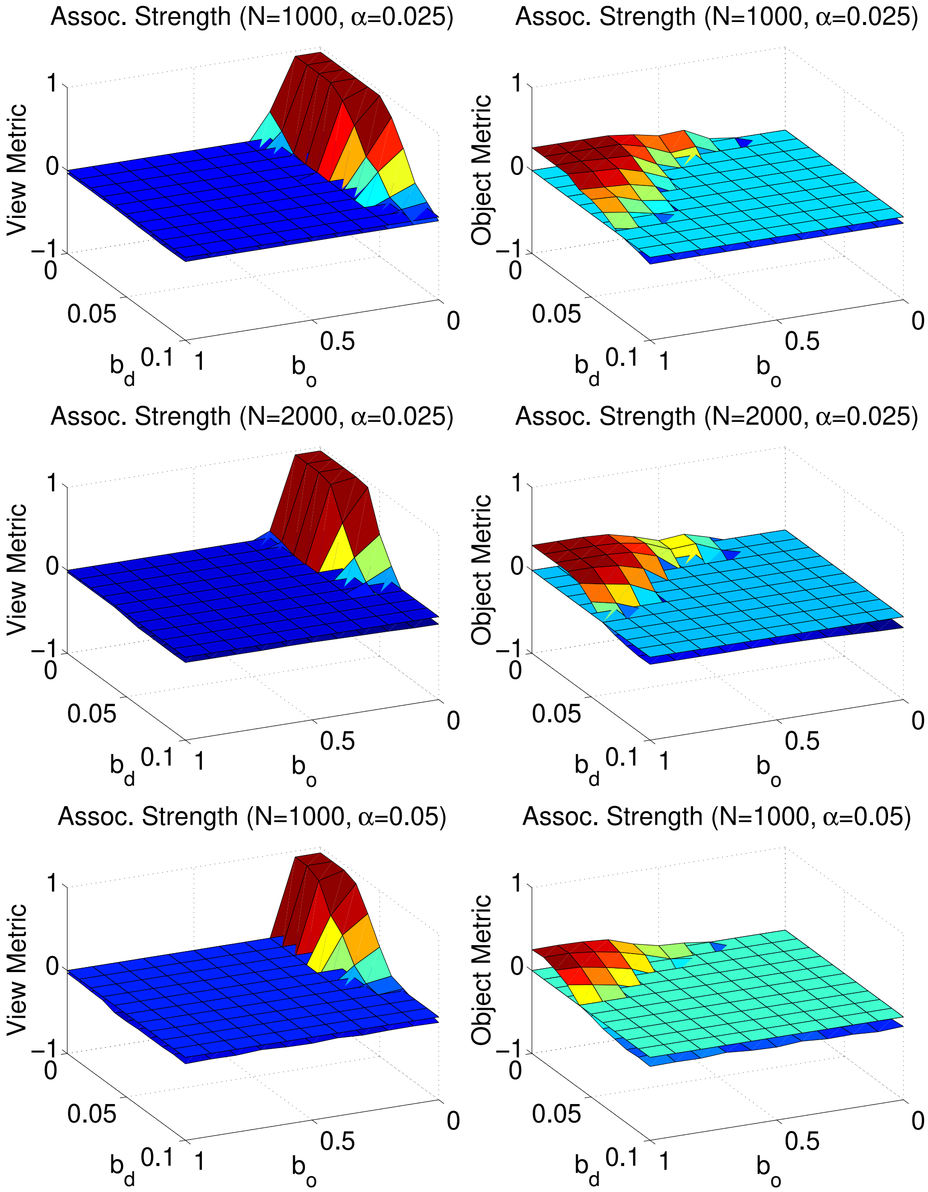

Fig. 3. Association Strength. View (left) and Object (right) metric performance given $d = 1$; $a = 0.5$ and $s = 5$ across ranges of association strengths b_o (between views of the same object) and b_d (between views of different objects). The top row shows results where $N = 1000$ and $\alpha = 0.025$; the middle row where $N = 2000$ and $\alpha = 0.025$; and the bottom row where $N = 1000$ and $\alpha = 0.05$.

finite size, with the networks sufficiently small (and lightly loaded) so that the destabilising term does not dominate the neuronal current.

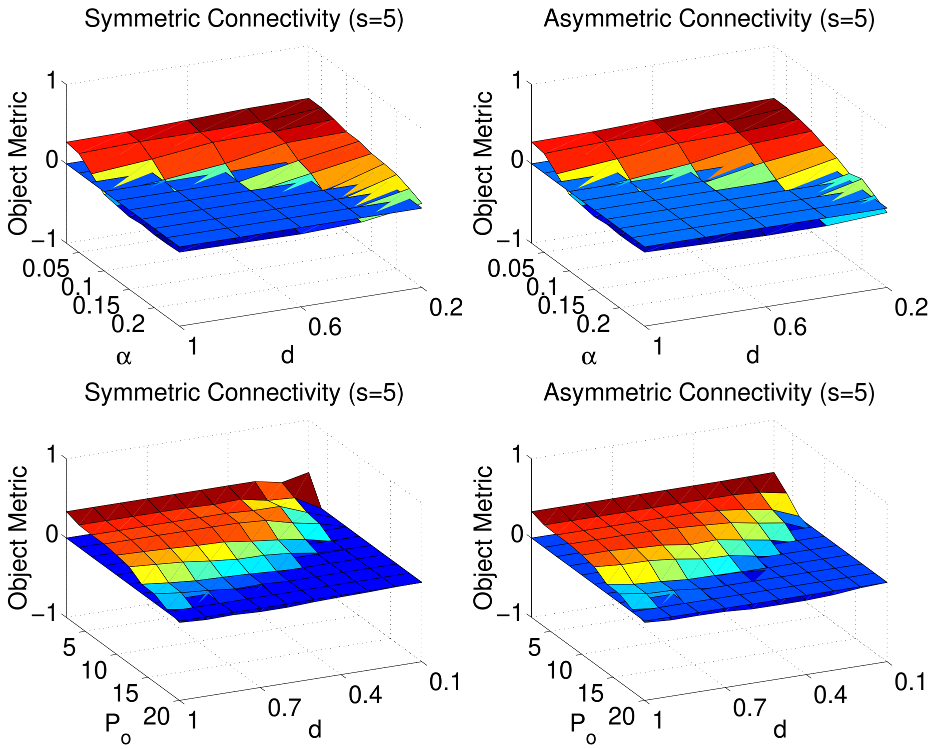

Fig. 4. Connectivity Dilution and Symmetry. Object metric performance given $N = 1000$; $b = 1$; $a = 0.5$; and $s = 5$ across a range of connectivity dilution (d) for both symmetric (left) and asymmetric (right) connectivity, presented given changing network loading (α, top) or the number of objects stored (P_o, bottom).

7 Connectivity Dilution and Symmetry

Connectivity dilution (d) refers to the probability that any two neurons are connected: $d = 0$ specifies no connectivity, while $d = 1$ specifies full connectivity. Symmetric connectivity includes the constraint that where connection J_{ij} exists (between neurons j and i), so connection J_{ji} (between neurons i and j) also exists, and is an important assumption underlying much analytical investigation (for the seminal example, see [12]). Asymmetric connectivity relaxes this constraint, such that there will not necessarily be a connection between two neurons in both directions. Asymmetric connectivity is of interest to explore by simulation, both because it is the usual situation in the brain, and because theoretical analyses usually assume symmetric connectivity (see [24]). We therefore describe the effects of dilution of both symmetric and asymmetric connectivity on the operation of this network.

Figure 4 shows performance detail given dilute connectivity in both symmetric (left) and asymmetric (right) form, presented based on changing network loading (α, top) or the number of objects stored (P_o, bottom). The number of views per object (s) was fixed at 5, and the strength of association between different views of the same object (b) was 1. Note that the results are shown over both α and P_o since the former allows comparison with the infinite-N analytical case, while the latter allows more detailed regular sampling (i.e. since $\alpha = sP_o/Nd$ is a function of d, and s and P_o are integers, so less dense *regular* sampling of $[\alpha, d]$-space is possible).

The first notable result is that performance under the symmetric and asymmetric paradigms is extremely similar: the graphs vary only in very minor detail. Second, performance generally decreases as either the network loading or the number of objects increases. Third, where the total number of objects (P_o) is kept fixed, so decreasing connectivity also decreases retrieval performance. However, where the number of objects per connection (α) is kept constant, decreasing the connectivity then *increases* retrieval performance: more dilute connectivity (i.e. fewer connections) allows better retrieval performance proportional to the overall number of connections available. Finally, for each number of objects stored (for a given network size and sparseness), there is a threshold level of connectivity which is required to gain maximal performance, and above which no additional connectivity provides additional performance.

8 Sparseness

Following Gardner [9]; Tsodyks and Feigel'man [32]; and others, we expected that decreasing pattern sparseness (a) would result in generally higher capacity (more patterns stored). Figure 2 shows object metric values using 5-view objects given sparseness values of $a = 0.5$ and $a = 0.1$, and is entirely consistent with this expectation. This section presents a systematic empirical investigation of the relationship between sparseness (a) and the number of views of each object (s).

Figure 5 shows values of the object metric for varying numbers of views per object ($s = 2$ through 7, each presented in a separate graph, with even values to the left and odd to the right); network loading (α); and sparseness (a). The maximal sparseness shown is 0.5: values greater than 0.5 give equivalent performance to values the same margin less than 0.5 (e.g. $a = 0.6$ yields equivalent performance to $a = 0.4$). The minimal sparseness shown is 0.1, with smaller values particularly susceptible to finite size effects ($N = 1000$). Finally, the strength of association between views of the same object (b) was fixed at 1. The notable results are as follows.

First, values of the object metric are *not* simply proportional to the decrease of sparseness, but instead vary gradually between performance peaks and troughs. Where performance is high (e.g. at the peaks), capacity *does* increase with decreasing sparseness, but the expected sparseness/capacity relationship

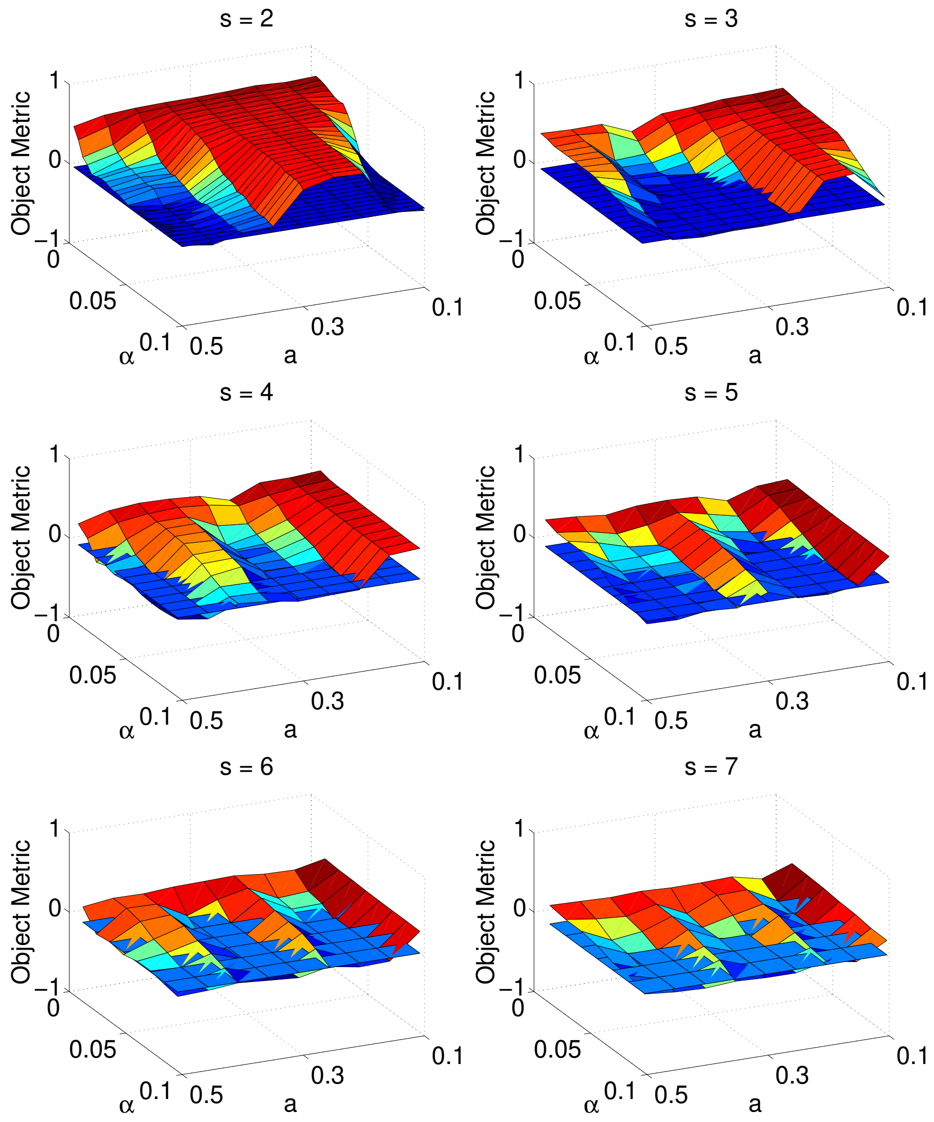

Fig. 5. Sparseness. Object metric performance given $N = 1000$; $d = 1$; $b = 1$; and $s = 2; 3; 4; 5; 6; 7$ across ranges of network loading (α) and sparseness (a).

does not hold in general.

Second, the number of performance peaks and troughs is directly proportional to the number of views per object (s). For example, $s = 2$ yields 1 peak at $a \approx 0.25$ and 1 trough at $a \approx 0.5$; $s = 3$ yields 2 peaks at $a \approx 0.17$ and 0.5, and 1 trough at $a \approx 0.33$; and $s = 4$ yields 2 peaks and 2 troughs.

Third, a sparseness of 0.5 (the left-hand edge of each graph) corresponds to either a performance peak or trough, depending on whether s is odd or even.

The third result is of particular interest here, since $a = 0.5$ was used in the analytical results of [16] and (consistent with Amit et al [2]) showed that the network has mixed symmetric states (i.e. with equal correlation with all s

patterns) only where s is odd. The empirical investigation has confirmed that analysis, showing that performance with even s and $a = 0.5$ is very poor[5]. However, figure 5 shows that with even s, performance is not poor at all values of a (e.g. given $s = 2$, there is a broad peak at $a \approx 0.25$), and likewise with odd s performance is not good at all a. Amit et al [2] show that the distribution of the stored patterns is a critical component underlying the existence of stable states, and thus we explore the relevant detail next.

Figure 6 shows the result of a statistical investigation into the "intra-object interference" between different views of the same object for the same parameter values as the simulations of figure 5. Such interference is expressed here in inverted form, being the probability that the absolute net effect on some synapse (J_{ij}) of storing *all* views of a single object is greater than some limiting value c (i.e. $P(|\Delta J_{ij}| > c)$). The actual limiting value used here is $c = 1/N$ since this yielded curves of a convenient level of detail: similar general forms of the curve result from different values of c, corresponding to the varying probabilities that the net effect of storing one object will become negligible given influences on this synapse from storing *other* objects. The graphs show two curves, with the solid line showing the theoretical statistical result (by calculating the probability and effect of the various possible pattern value combinations at neurons i and j) at a resolution of $\Delta a = 0.01$, and a dotted line connecting the empirical result given by storing a single object in the model itself (averaged over 8 different random seeds, shown between standard error bars). The form of these curves matches the simulation performance curves in great detail, with the number and location (in a) of performance peaks and troughs all but identical.

The rationale behind this statistical approach is that where an even number of influences exist, equal but opposite influences can effectively cancel each other: at any one synapse, the individual effect of storing any one view may be substantial, but the net effect of storing several views may be nothing at all. As one changes sparseness (a) and the number of views per object (s), so the probability of balancing influences also changes, and gives rise to the

[5] The non-zero object metric at $a = 0.5$; even s; and very low loadings is a very special case. The typical scenario includes an attractor which is not equally correlated with all object views. Such *asymmetric* attractors (which are not accommodated by the underlying theoretical framework) do gain a positive object metric, since the attractor is more correlated with all views of this object than with any view of any other object. The object metric could be modified to detect such attractors, returning zero if the difference between the minimum and maximum correlations between the attractor and different views of the cued object was greater than a theoretical limit ($\theta = O(\sqrt{1/N})$). Empirical analysis using such a metric reduced quantitative performance values given even s and $a = 0.5$, but otherwise made no difference to the qualitative results presented. Note that the empirical finite size environment makes truly symmetric correlations most unlikely.

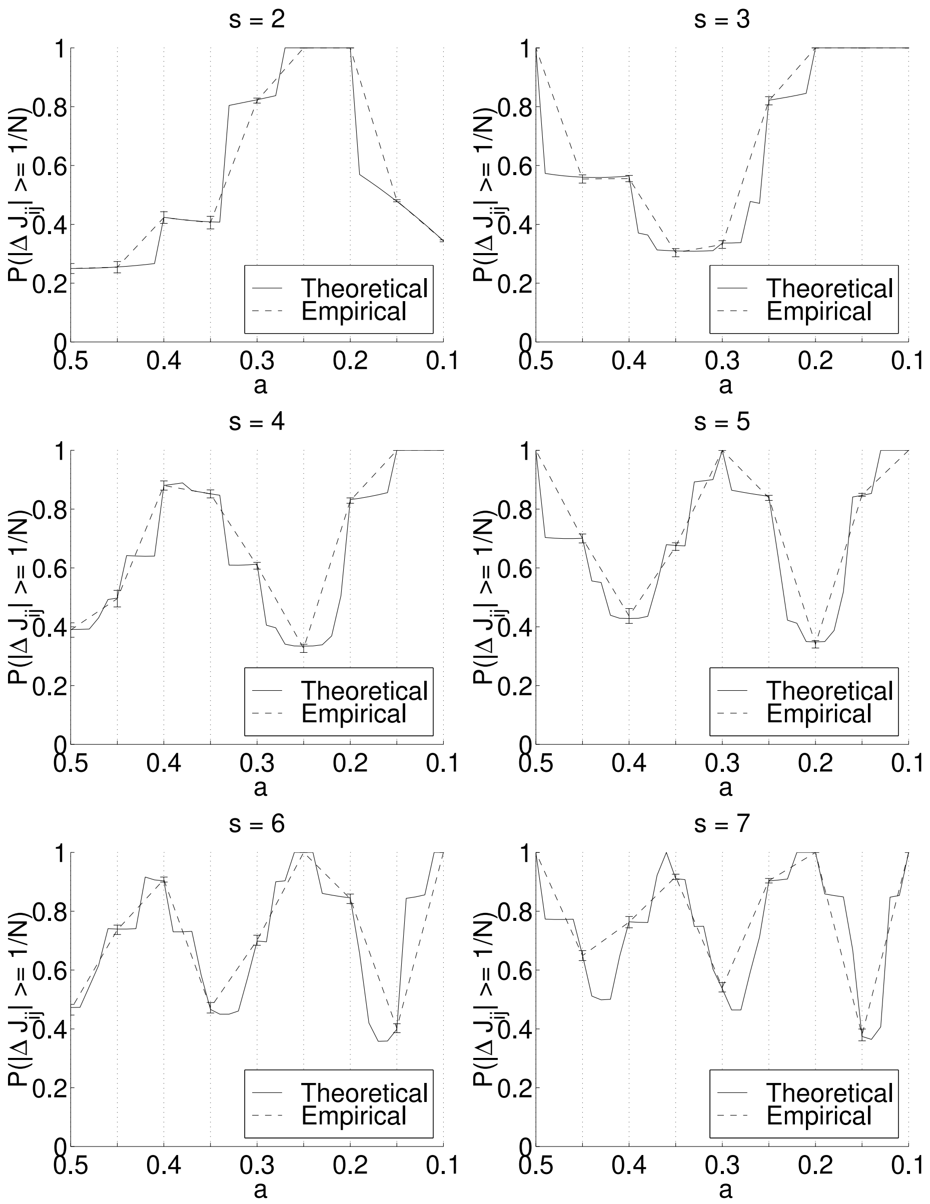

Fig. 6. Intra-object interference probabilities given $N = 1000$; $d = 1$; $b = 1$ and across a range of sparseness (a), when storing a single object of $s = 2; 3; 4; 5; 6$; or 7 views (i.e. $\alpha = 0.002; 0.003; 0.004; 0.005; 0.006$; and 0.007 respectively).

performance variation shown. Appendix B provides a signal-to-noise analysis consistent with the empirical results presented.

Note that the intra-object interference effect varies not only with sparseness (a) but also with the strength of association between different views of the same object (b). Figure 7 shows the quantitative interference probabilities given $s = 5$; $a = 0.5$; and b in the range [0,1]. The probability of avoiding interference is maximal (1) at the limits of b, and minimal (0.71) at values of b in the range [0.4002,0.4997].

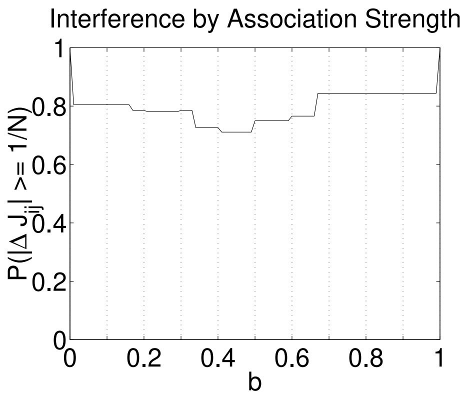

Fig. 7. Intra-object interference probabilities given $N = 1000$; $d = 1$; and $a = 0.5$ for varying b when storing a single object of $s = 5$ views (i.e. $\alpha = 0.05$).

9 Discussion

The results presented in this paper are consistent with and support the analytic model introduced by Parga and Rolls [16], showing how an attractor network can retrieve an object-specific representation given presentation of any view of that object. The application we have in mind is invariant object recognition, although clearly the results are quite general: each "object" could be considered an arbitrary pattern class, with each "view" an arbitrary member of that class. Thus, while we have presented the results in terms of recognition which is invariant across object views (i.e. rotation in depth), the approach is of course relevant to other types of invariant recognition, including position invariance; size invariance; and rotation invariance. Wallis and Rolls ([34]) investigated a feed-forward architecture which also exhibited invariance, with a particular utility of the current approach being its analytic tractability: issues such as network capacity can be analysed and measured quantitatively. The empirical results presented here show good agreement with the analytic prediction.

The analytic model introduced a specific issue: at sparseness (a) of 0.5, stable symmetric states exist for odd numbers of patterns (s), but not for even. The results described in this paper show that this result holds empirically (see figure 5, where the object metrics at $a = 0.5$ show a peak given odd s but a trough given even s) , but also that the same approach to transform invariance does hold for even numbers of views of each object: all that is required is an appropriate level of pattern sparseness. The relationship between pattern sparseness and object-retrieval performance is entirely predictable for a given number of views per object (see figure 6), being a simple consequence of the probability of binary patterns to cause interference by cancellation. The specific strength of the association between different views of the same object (b) also affects the degree of such interference, although the effect is weak. Reduction of the general effect of interference by cancellation is likely to involve systems which can store non-binary patterns, thus including neurons which can assume non-binary firing rates, and will be the subject of a future investigation.

The results described here also extended the previous analytic results in a number of other ways, and are of particular relevance to how such networks might operate in the brain.

First, an important issue is that the approach was shown to work well not only with diluted connectivity, but also with asymmetric connectivity (see figure 4). The assumption of symmetric connectivity which underlies the analytic treatment is shown here to be unnecessary for effective network operation.

Second, as expected, reducing the sparseness increased the numbers of different objects that could be stored in the network (see figures 4 and 5), albeit that it is not independent of the cancellation effect described above (i.e. the increase is seen only by comparing capacity at performance peaks). Reducing the sparseness is effective of course only within limits, determined in practice largely by finite size effects that occur in simulations but which would be less likely to occur in the brain given the large numbers of neurons and synapses involved in typical networks (see, for example, [24]).

Third, the empirical analysis also showed the model's tolerance to cue distortion, showing good retrieval of objects even when the cue was distorted (see figure 2). Such cue distortion is relevant to normal invariant pattern recognition, in that we rarely see a retrieval cue, such as a view of an object, in exactly the same way that we originally learned about that object. The network in this sense could complete from incomplete views, and generalise from similar views.

Fourth, the simulations also showed that the system would still operate correctly when there was some association between the views of different objects. This is relevant to systems which learn invariant representations by incremental means, for example by using a trace rule ([8,17,10,34] c.f. [26]). In such a scheme, if the decaying memory trace which enables association between different views of an object is not reset between objects, there will be some (though weak) association between views of different objects. The results described here show quantitatively how a system of finite size operates under conditions when such associations exist. These associations must of course be much smaller than the associations between the different views of a single object.

Fifth, the investigations described here also provide an explanation of why the results of Parga and Rolls (1998) could only be applied directly when odd numbers of views of each object were stored. It was shown here that peaks and troughs in the performance graphs occurred depending on the number of views (s) of each object that were being stored, and the sparseness a of the representation. These peaks and troughs were shown to be related just to whether or not with the binary patterns used there were particularly large values of synaptic weights which resulted from interference between the patterns. The results of Parga and Rolls (1998) involved investigations only with a=0.5, and with binary patterns and even numbers of views of each object, interference at synapses occurred, and correct storage could not be demonstrated, although it could if odd numbers of views of each object were stored. As shown here, this interference is not a fundamental limitation of the model at all, but is just due to the particular way in which binary patterns can interact. In the brain, where the storage would not be of strict binary numbers but something which more closely approximates continuous firing rates of neurons to different associated views, and where the number of synapses per neuron would be much larger, the variations in storage ability associated with statistical fluctuations of binary combinations would not apply, and the system would be expected to perform consistently well regardless of the particular number of views of each object stored.

In conclusion, the work described in this paper has shown by simulation that in an attractor network in which the storage capacity can be rigorously investigated, the network can perform well at invariant object recognition when the different exemplars of each object are associated together in a synaptic matrix of the type that could be set up by the operation of a trace learning rule (see Wallis and Rolls [34] and Rolls and Milward [21]).

A Pattern Domain

Identical results to the above can be obtained with an arbitrary choice of binary pattern values: the [0,1] language used above was purely for notational convenience, with [-1,+1] or [12,93] or whatever able to be used without incident. The necessary changes to the above formalism are, first, to change the pattern generation, distortion, and translation formulae so that arbitrary minimum (ξ_{min}) and maximum (ξ_{max}) token values are used rather than the implied 0 and 1. The new translation function, for example, becomes :-

$$g(I) = \begin{cases} -1 \ I = \xi_{min} \\ +1 \ I = \xi_{max} \end{cases} \tag{A.1}$$

Second, raw a is no longer the pattern mean (ξ_μ), and thus the synaptic matrix calculation becomes :-

$$\mathcal{J}_{ij} = \frac{d_{ij}}{\sigma_\xi^2 N d} \sum_{\beta=1}^{P_o} \sum_{\mu,\nu=1}^{S} (\eta_i^{\beta\mu} - \xi_\mu) X^{\mu\nu} (\eta_j^{\beta\nu} - \xi_\mu) \tag{A.2}$$

where σ_ξ^2 is the pattern variance.

Note that the effect of these changes is to make explicit the separation between the particular values chosen for the pattern domain and the actual neuronal outputs ($\mathcal{V}$) as prescribed by Φ, thence building an identical synaptic matrix independent of the specific ξ_{min} and ξ_{max} chosen. Since these domains may well be different ($g()$ may not be the identity function), performance evaluation throughout this work is based on state/pattern *correlation* rather than simple overlap.

B Intra-Object Interference Analysis

An alternative explanation of the destabilisation of the object phase for some values of sparseness a and number of object views s can be given by a "signal-to-noise" analysis. Such an analysis studies the relative contribution of the object-selective (signal) and object-unselective (noise) components of the effective current received by a given neuron when the network is in a particular state. Note that our primary interest is in the stability of the object phase, and thus we here perform the signal-to-noise analysis assuming that the network is in a symmetric object attractor state (i.e. with equal overlaps with all views of a given object). Further, for simplicity, we also assume that the network is fully connected ($\forall_{i \neq j}\ i, j : d_{ij} \equiv 1$), although the result can easily be extended to accommodate dilute connectivity.

The effective current (or local field) to neuron i given network state $\{\nu_j\}$ (where $j = 1, \ldots, N$) is :-

$$h_i = \sum_{j \neq i}^{N} \mathcal{J}_{ij} \nu_j = \frac{1}{Na(1-a)} \sum_{j \neq i}^{N} \sum_{\beta}^{P_0} \sum_{\mu,\nu}^{s} \tilde{\eta}_i^{\beta\mu} \chi^{\mu\nu} \tilde{\eta}_j^{\beta\nu} \nu_j \tag{B.1}$$

using $\tilde{\eta}_i^{\beta\mu} \equiv (\eta_i^{\beta\mu} - a)$.

To simplify the analysis, we change network state representation: individual neuronal activities (-1,+1) become $(0, 1)$ by $\nu_j = 2\nu_j^{'} - 1$. We then separate object β_0's contribution $S_i^{\beta_0}$ from R, the sum of the contribution of all objects different from β_0. The signal and noise components of the effective current are $S_i^{\beta_0}$ and R respectively, given by :-

$$h_i = \frac{2}{Na(1-a)} \sum_{j \neq i}^{N} \sum_{\mu,\nu}^{s} \tilde{\eta}_i^{\beta_0 \mu} \chi^{\mu\nu} \tilde{\eta}_j^{\beta_0 \nu} \nu_j^{'} \quad + \quad R \quad \equiv \quad S_i^{\beta_0} \quad + \quad R \tag{B.2}$$

Note that the representation change means that the original signal is effectively decomposed into two components: one proportional to ν', represented by $S_i^{\beta_0}$; and one independent of ν, and with zero mean, which is thus included in R.

The magnitude of R is a random quantity with mean zero and variance proportional to network loading α, and is independent of both the object-specific signal ($S_i^{\beta_0}$) and the neuron (i) in which the current applies. For our purposes here it is sufficient simply to note that R causes a random fluctuation of order α in the effective current to any neuron.

Dropping the object index from $S_i^{\beta_0}$, and defining the overlap of the network state with a given view as :-

$$m^{\mu} \equiv \frac{1}{Na(1-a)} \sum_{j}^{N} \tilde{\eta}_j^{\mu} \nu_j^{'} \tag{B.3}$$

the signal can be expressed as :-

$$S_i = 2 \sum_{\mu,\nu}^{s} \tilde{\eta}_i^{\mu} \chi^{\mu\nu} m^{\nu} \tag{B.4}$$

We can then use the assumption of object attractor symmetry (i.e. $\forall \mu : m^{\mu} = m$) to obtain :-

$$S = 2m[1 + b(s-1)] \sum_{\mu}^{s} \tilde{\eta}_i^{\mu} \equiv m[1 + b(s-1)]z \tag{B.5}$$

where we have defined :-

$$z = \sum_{\mu}^{s} \tilde{\eta}_i^{\mu} \tag{B.6}$$

The signal at neuron i given that the network is in an object attractor state is, therefore, proportional to the sum of the differences between the activity value of neuron i in all views of that object and the neuron's mean activity. As we will now show, given the stochastic nature of the patterns which represent the views, there are certain values of the parameters s and a for which there is a non-zero chance that this sum (the variable z) will vanish. Where z is zero, the receiving neuron will be driven entirely by the noise component of the current: the subsequent state will be random, and thus any object state would be destabilised.

The explanation requires the probability that z vanishes, a value which depends on the probability distribution of the η_i. This distribution can be interpreted as the fraction of neurons which receive zero signal, able to be calculated as follows.

The variable z can take $s+1$ different values, depending on the number k of views in which neuron i takes the value 1. One can use this number to parametrise z, since $z \equiv z_k = k - sa$ where $k = 0, \ldots, s$. One immediately sees that if $n = sa$ is an integer, z will vanish at one or more non-zero values of k (i.e. where $k = 1, \ldots, n$). The probability distribution of z_k is simply :-

$$P(z_k = k - sa) = a^k (1-a)^{s-k} \frac{s!}{k!(s-k)!} \tag{B.7}$$

and thus, for given (a, s) such that $sa = n$ is an integer, the fraction p_R of neurons driven solely by noise is :-

$$p_R = \sum_{k=1}^{n} P(z_k) \tag{B.8}$$

As an example, if $a = 0.5$ and $s = 4$ so $p_R = \frac{5}{8}$. This means that if the network were originally in an object state, the very next iteration would result in more than half of the neurons updating according to noise R alone. Since this noise is uncorrelated with the signal, an average fraction of $p_R/2$ neurons will update incorrectly, making this state highly unstable.

This analytical result agrees with the numerical results of figure 5, such that for each number of views per object s, the values of a which give a performance minimum are integer multiples of $1/s$.

References

[1] D. J. Amit. *Modelling Brain Function.* Cambridge University Press, 1989.

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985.

[3] M. S. Bartlett and T. J. Sejnowski. Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network: Computation in Neural Systems*, 9(3):399–417, 1998.

[4] E. Bienenstock and C. von der Malsburg. A neural network for invariant pattern recognition. *Europhys. Lett.*, 4:121–126, 1987.

[5] M. C. A. Booth and E. T. Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 8(6):510–523, 1998.

[6] J. Buhmann, R. Divko, and K. Shulten. Associative memory with high information content. *Physical Review A*, 39(5):2689–2692, 1989.

[7] V.S. Dotsenko. Neural networks: translation-, rotation-, and scale-invariant pattern recognition. *J. Phys. A*, 21:L783–L787, 1988.

[8] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

[9] E. Gardner. Maximum storage capacity in neural networks. *Europhysics Letters*, 4(4):481–485, 1987.

[10] M. Griniasty, M. Tsodyks, and D. Amit. Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation*, 5(1):1–17, 1993.

[11] C. G. Gross, R. Desimone, T. D. Albright, and E. L. Schwartz. Inferior temporal cortex and pattern recognition. *Experimental Brain Research Supplement*, 11:179–201, 1985.

[12] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA*, 79(8):2554–2558, 1982.

[13] Y. Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817–820, 1988.

[14] Y. Miyashita and H. S. Chang. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, 331(6151):68–70, 1988.

[15] R. C. O'Reilly and M. H. Johnson. Object recognition and sensitive periods: a computational analysis of imprinting. *Neural Computation*, 6(3):357–389, 1994.

[16] N. Parga and E. T. Rolls. Transform-invariant recognition by association in a recurrent network. *Neural Computation*, 10(6):1507–1525, 1998.

[17] E. T. Rolls. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosphical Transactions of the Royal Society of London B: Biological Sciences*, 335(1273):11–21, 1992.

[18] E. T. Rolls. Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes*, 33(1–2):113–138, 1994.

[19] E. T. Rolls. Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research*, 66(1–2):177–185, 1995.

[20] E. T. Rolls. A neurophysiological and computational approach to the functions of the temporal lobe cortical visual areas in invariant object recognition. In M. Jenkin and L. Harris, editors, *Computational and Psychophysical Mechanisms of Visual Coding*, pages 184–220. Cambridge University Press, Cambridge, 1997.

[21] E. T. Rolls and T. J. Milward. A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation*, page in press, 1999.

[22] E. T. Rolls and M. J. Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, 73(2):713–726, 1995.

[23] E. T. Rolls and A. Treves. The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network: Computation in Neural Systems*, 1(4):407–421, 1990.

[24] E. T. Rolls and A. Treves. *Neural Networks and Brain Function.* Oxford University Press, 1998.

[25] K. Sakai and Y. Miyashita. Neural organisation for the long-term memory of paired associates. *Nature*, 354(6349):152–155, 1991.

[26] R. S. Sutton and A. G. Barto. Towards a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, 88(2):135–170, 1981.

[27] K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66(1):170–189, 1991.

[28] A. Treves. Graded-response neurons and information encodings in auto-associative memories. *Physical Review A*, 42(4):2418–2430, 1990.

[29] A. Treves. Are spin-glass effects relevant to understanding realistic auto-associative networks. *Journal of Physics A: Mathematical and General*, 24(11):2645–2654, 1991.

[30] A. Treves. Dilution and sparse coding in threshold-linear nets. *Journal of Physics A: Mathematical and General*, 24(1):327–335, 1991.

[31] A. Treves and E. T. Rolls. What determines the capacity of auto-associative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397, 1991.

[32] M. V. Tsodyks and M. V. Feigel'man. The enhanced storage capacity in neural networks with low-level activity. *Europhysics Letters*, 6(2):101–105, 1988.

[33] C. von der Malsburg and E. Bienenstock. A neural network for the retrieval of superimposed connection patterns. *Europhys. Lett.*, 3:1243–1249, 1987.

[34] G. M. Wallis and E. T. Rolls. Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194, 1997.

[35] V. Yakovlev, S. Fusi, E. Berman, and E. Zohary. Inter-trial neuronal activity in inferotemporal cortex: A putative vehicle to generate long term visual associations. *Nature Neuroscience*, 1(4):310–317, 1998.