Recognition of Objects Rotated in Depth Using Partial Synchronization of Chaotic Units

David L. DeMaris
University of Texas at Austin
Dept. of Electrical and Computer Engineering
demaris@ece.utexas.edu

Baxter F. Womack
University of Texas at Austin
Dept. of Electrical and Computer Engineering
womack@ece.utexas.edu

Abstract

A regular array of discrete-time nonlinear oscillators with recurrent connections (coupled map lattice or CML) can perform object recognition by acting as a dynamical recognizer, while simultaneously performing computations to normalize class members to a common representation. Partition cells of the network state space serve as dimensions of the representation space. Occupancy statistics in each such cell after a brief evolution of the system, governed by intrinsic time-varying dynamics of the oscillators and the forms presented as initial conditions, result in a population code measured over all units. Results on recognition of paperclip objects rotated in depth using an ensemble of classifiers are reported. With training on seven views separated by 30 degrees, the system achieves recognition rates of 85% for a set of twenty paperclip objects. Recognition is achieved in an average of 12 iterations of the recurrent system. Performance degrades with fewer training views, with angular distance from training views, and with an increasing number of objects. Biological support for this theory of object recognition is examined.

1 Introduction

While the nature of object representation in the human and other primate brains remains controversial, progress on understanding the nature of internal representations has been made in recent years. This progress rests largely on observing indirect measures such as error rates in human and primate recognition tasks, as well as neural spike rate recordings in single and multi-channel paradigms. Prediction of a behaviorally correlated stimulus correlated with behavior from the neural trace is the most informative neural analysis technique.

Theories on object recognition classically fall into two major camps: *structural* or *object based* theories, and *view based* theories. Structural description theories assume that *view-independent* or i*nvariant* features underlie the representation of objects. View based theories build representations from various local features extracted from separate learned views, and have historically been associated with the subordinate level of categorization. Advocates of *view-based representations* have recently claimed that basic level categorization can emerge in a natural fashion from the clustering involved in making subordinate level distinctions [1]. A growing body of experimental evidence now suggests that performance on recognition tasks is proportional to the distance from the nearest familiar view. This includes both error rate measures [2] and recognition time [3][4].

In everyday life we spontaneously identify objects, and create categories from the diverse retinal images of an object seen from different viewpoints, or from the diverse individual members of a species. However, this fundamental problem of stimulus identity has only recently begun to be addressed satisfactorily in computer vision. While many geometric methods have been developed for handling translation and rotation in the plane [5][6], rotation *in depth* of diverse objects has been addressed most recently and successfully through statistical approaches with a rich feature space [7], and by an ensemble of radial basis function (RBF) neural networks implementing a view normalization and interpolation strategy [8].

While the RBF method has produced impressive recognition rates on small object worlds, the fundamental computational and coding principles (integration of spike rates and rate codes transmitted by specific neurons) on which its status as a biological theory rests have been questioned recently in the neuroscience community [9]. The work presented here rests on a different view of neuronal function in higher level vision in which neurons perform the work of perception chiefly through participation in medium and large scale oscillating ensembles. These oscillations are aperiodic due to the underlying nonlinear dynamics of such ensemble units, but strong coupling can suppress the nonlinearity or chaos level, resulting in synchronized chaos, clusters of synchronized chaos, or even more regular phenomena such synchronized periodic oscillations, perhaps with instabilities. The underlying computational principles for higher spatial vision are the interaction of nonlinearity and synchronization phenomena with oriented coupling kernels.

An additional important aspect of the network described here is the initial confinement of the network state to a subspace (by the extraction of a primal sketch in early vision), which then undergoes phase space expansion by the chaotic dynamics, and ultimately reaches a decision state in another subspace. Thus two related types of synchronization are operating: the usual sense of phase synchronization, and the more abstract sense from graph theory and symbolic dynamics which we designate "subspace synchronization".

Statistical population coding is used, meaning that statistics over the entire population of units are measured; objects are located as a *subspace* in the space whose dimensions are defined by intervals in the dynamical phase space of the system. This is in contrast with gnostic cell or place coding formulations, or distributed coding in connectionist networks as activation of specific units. Instead, partial phase synchronization after some iterations results in what might be called *transient subspace synchronization*. In the latter, the effective correlations between units are increased, and effective dimension (phase space volume) is reduced.

These principles are motivated and supported by a variety of observations in primary visual areas and various sub-regions of inferotemporal cortex implicated in object recognition; for reviews see [10],[11],[12] In addition, the work is compatible with recent approaches in computer vision which have been designated as "shape through transformation" [13].

While building on past work on recurrent networks and high dimensional coupled chaotic systems, this study makes substantive breaks with previous approaches. The most important is to introduce two stages with a sharp change in the parameters, rather than stationary or smoothly changing dynamics. It also breaks with typical recurrent networks by concentrating on the transient regime of dynamics rather than on attractors. This staged processing, with desynchronization and partial synchronization of dynamical transients, motivates our designation of the system as a *Synchronization Opponent Cooperative Activity* (**Soca**) network.

2 Methods

In this section we present the basic formalism of a coupled map lattice [14], followed by the extension to a time varying system. Input images in the system are subjected to a thresholding operation to produce a silhouette. This binary image is mapped into values 0.0001 and .9999 as an initial condition which is transformed by the coupled map lattice.

2.1 Coupled Map Formulation

The coupled map computation consists of interleaved diffusive coupling and nonlinear map stages proceeding for some number if iterations or time steps t. The first diffusive coupling step is expressed as:

$$S_d(x,y) = (1-c)S_t(x,y) + \frac{c}{4}\left[S_t(x,y+1) + S_t(x,y-1) + S_t(x+1,y) + S_t(x-1,y)\right] \quad (1)$$

where: d is the intermediate diffusion array, t is the current time step, x, y are the spatial indices of the pixel array S at the center of the diffusion neighborhood, S is the state variable at each site of the shape array, and c is the coupling constant restricted to the range <0.0 to 1.0>. The fact that diagonal elements not processed results in an orientation sensitivity of the process.

The second computational unit applied in each time step is the asymmetric logistic map:

$$S_{t+1} = 1 - bS_t^2 \ , \ \begin{cases} -1.0 < S < 1.0 \\ 0.0 < b < 2.0 \end{cases} \quad (2)$$

where b is a bifurcation parameter; changing this parameter forces a structured transition between phases following the sequence of attractor types: fixed point → limit cycle cascade of increasing period and instability → intermittency → chaos → {limit cycle cascade →chaos}. In contrast to the symmetric logistic, this function ranges from –1.0 to 1.0, and is thus suitable for modeling neural dynamics which vary about a background rate.

2.2 Synchronization Opponent System

The two stages just described constitute a single iteration of the CML system. Motivated to investigate slower dynamics as modulating bifurcation and coupling parameters to achieve some computational goal, we utilize two stages designated as synchronization opponents. The first is desynchronizing (both phase and subspace senses), while the second is synchronizing. Each stage, or *attractor frame* consists of a triple {b, c, s}, where s is number of iterations in the stage. The terminology attractor frame is introduced to clarify the point that the computation is performed in the transient evolution, rather than by reaching attractors.

Two such frames are applied in turn. During evolutionary search for the parameter sets, the first stage is constrained to 2-6 iterations, the second constrained to 2-9 stages. In the evolved solutions, the total iterations required to create the representation space ranges from 6 to 14 iterations, with a mean of 11.3 iterations for the 39 objects in a 3D object recognition task. The following pseudocode summarizes the procedure.

```
procedure  synchronizationOpponentNetwork
  image = threshold(downSample(readImage)))
  for iterations = 1 to t1
          diffuseImage = filter2D(couplingMatrix,image)
          image = logisticMap(diffuseImage,b1,c1);
  end // Desynchronize stage
  for iterations = 1 to t2
     diffuseImage = filter2D(CouplingMatrix,image)
     image = logisticMap(diffuseImage,b2,c2);
  end // Partial Synchronize stage
end  // procedure
```

Because of the variation in divergence and convergence times, a specific set of bifurcation, coupling, and iteration time parameters {b1, c1, t1, b2, c2, t2} has a characteristic response to any given image or family of images. Each image can be considered as a set of overlapping initial configurations of size t1+t2; by the end of the Soca process above, information about local configurations from a window of size t1+t2 is contained in each unit (i.e. each pixel in the processing array). The set of initial configurations comprising one image may be highly synchronizing for those parameters, while those from another image may be less so.

The intuition behind the network operation is that images in some category are considered as *productions* of a *stochastic language* on an alphabet α whose symbols are local pixel configurations. We seek parameters for the first (desynchronizing) stage which, for this language, have the appropriate divergence rate matching parameter determined characteristics of the second (synchronizing) stage. The second stage must have characteristically avoided regions of state space and state transitions such that images in the category will converge near a characteristic sparse distribution or subspace.

2.3 Dynamical Recognizers for View Normalization

One way to view a family of images (such as outline views of an object) is as a stochastic language, where the alphabet is particular orientation and transitions probabilities of various orders capture the adjacency or co-occurrence of features. The problem of learning to accept positive exemplars of a language while rejecting negative exemplars is known as language induction. Classical machine learning approaches to this problem construct a finite state automaton to affect recognition. Formally, a finite state recognizer is a quadruple $\{Q, \Sigma, \delta, F\}$, where Q is a set of states (with q_0 denoting the initial state), Σ is some finite alphabet, δ is a transition function mapping $Q \times \Sigma \Rightarrow Q$, and $F \subset Q$ is a set of final or accepting states. A string of tokens from alphabet Σ is *accepted* by the recognizers if, starting from initial state q_0 the sequence of state transitions indicated by the tokens in the string ends up in one of the final states in subset *F*.

The dynamical recognizer [15] is a quadruple $\{Z, \Sigma, \Omega, G\}$, where $Z \subset R^k$ is a state space; $z_k(0)$ is the initial condition. Σ is the input "alphabet", where a particular closed interval in Z corresponds to each element in this alphabet. . Ω is the dynamic, a sequence of transformations $\omega_l{:}Z \rightarrow Z$ (one for each token) with an associated set of dynamical parameters; these parameters are fixed for a particular recognizer during the induction (training) process. $G(Z) \rightarrow \{0,1\}$ is the decision function which maps one or more states in the sequence produced by the dynamic to an accept/ reject decision. In Pollack's intial study cited above, only the final state and token are used in the decision function.

The recognition method here is consistent with and extends the dynamical recognizer framework to higher dimensional systems. A key difference is that the Soca net operates on an image "string" in parallel (thus the network state space has dimensionality R^N, where N is the number of pixels or sampled image elements), and the tokens are used only once as the initial state. Such a parallel recognizer framework for *picture languages* with constrained states and transformations was studied in a series of papers by Rosenfeld [16]. In Rosenfeld's formulation, the transition function at each pixel is now a function of several tokens in some spatial neighborhood, i.e. a cellular automata formalism. The decision function is necessarily modified by this larger state space. Rosenfeld proposed several possibilities:

1. every spatial element reaches an accepting state;
2. any element reaches an accepting state;
3. one particular spatial element reaches an accepting state;

In the Soca network and recognition strategy, we form a representation space, but that space consists of *statistics measured instantaneously during a high dimensional, parallel dynamics*, rather than a direct map of the input features or measurement space. These statistics naturally support a decision function; one simply defines a threshold distance for each classifier, and accepts an object as an instance of language L if the Euclidean distance from the classifier's occupancy statistic signature to the corresponding learned signature is within the threshold. The distances might vary by class, depending on the cluster density of that class in the representation space. In practice, the present network treats the partition cells occupancies of all classifiers as a representation space of a quasi-metric character, as objects whose orientation statistics are similar will map to nearby points in the representation space. We call the space a *partition cell metric space*. Thus any presented outline image has a computable distance function to each recognizer.

2.4 Learning

To produce such a recognizer which can capture the differences in orientation statistics in the family of images, we employ an evolutionary computing strategy. In contrast to typical practice, the generated network is not tested on the recognition task, but is evaluated on several abstract measures on the partition cell distribution in an objective function.

The CML computation is relatively expensive on serial hardware. In the present recognition implementation each classifier previously generated for different objects would be evaluated during evaluation of a candidate network genotype. Using Matlab, we presently evaluate 3000 candidate networks (30 generations of 100 individuals) in about 45 minutes per object on a 400MHz PowerPC 750; full evaluation on the recognition task would increase the time by a factor of 200. Evaluation of a single parameter set on 75 x 75 pixel test images uses at most .18 ms, so recognition performance approaches real time human performance for 20 objects.

The objective function used in learning must balance two goals: 1) tight clustering of different views of an object and 2) separation of classes corresponding to various views of other objects. A view normalization strategy (originated in the radial basis function approach to view based recognition) serves the tight clustering goal, with a cross-entropy term comparing current output with previously stored representations enforcing separation of objects in the representation space.

The objective function returning a fitness f, with low values indicating higher fitness, takes the form

$$f = W_d D - W_e(H_c + H_s) + P_{synch} \tag{3}$$

where W_d is the inter-view distance or normalization weight and W_e is the entropy weight; D, the inter-view distance sum for j views is

$$D = \sum_{ij} \sqrt{\left(\sum_{p=1}^{k} v_{i,p} - v_{j,p} \right)} \tag{4}$$

where v_p are the occupancies of k partition cells for each of the j views.

H_c is the cross entropy or Kullback-Lieber information measure between the current reference view distribution C and the database of N object distributions (signatures) S with k bins:

$$H_c = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} C_j \log_2 \frac{C_j}{S_{i,j}} \tag{5}$$

H_s, the Shannon entropy of the current signature with k bins is

$$H_s = \sum_{i=1}^{k} S_i \log_2 S_i \tag{6}$$

and

$$p_{synch} = \begin{cases} 1000 \text{ if } \max(S_i) < \text{synchPenThresh} \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

The parameter synchPenThresh was empirically determined as .15. The following table indicates results of varying the synchronization penalty threshold on the nearest neighbor match of the 0° view.

Table 1. Error rates for objective function weight sets

W_d	We	SynchPenThresh	Error Rate %
20	2	.4	45
20	2	,25	24
20	**2**	**.15**	**15**
min	Max	.15	40

The table above is a comparison of effects of different sets of objective function parameter weight on performance of nearest neighbor match with 20 objects and training with all views. The last row indicates that rather than two terms with weights, the objective was formulated as a minimizing the *ratio* distance / entropy. The set W_d=20,W_e=2, SynchPenThresh =.15 was selected as the *standard weight set* for learning and recognition trials.

2.5 Recognition

Having generated representations by searching for dynamics which perform both view normalization and class separation in the space via a cross entropy term, recognition of the objects is implemented by simply processing a target view with the parameter set {*b1,c1,t1,b1,c2,t2*} defining each classifier, then computing the Euclidean distance to a saved histogram or distribution obtained by averaging each partition cell occupancy value across all views of the original object. The number of cells (64 in the current implementation) is an implicit parameter. The object is assigned to the class which results in the minimum distance to the original signature.

3 Data Selection

Much recent experimental visual psychology and neuroscience research on the on 3-D object recognition has worked with a family of objects commonly referred to as paper clips. For continuity with this research, we performed network simulations on a set of paperclip objects previously studied in psychophysics. The image set designed in the visual psychology lab of M. Tarr, and consists of 39[1] *paper clip* objects, with seven views provided for each object rotated in depth. Each object is a chain of 5 cylinders, with a variable joint angle connecting each pair. The views are separated by 30°, ranging from –90° to 90°. The objects were originally developed to answer questions regarding the *recognition by components* or *geon* theory of Biederman [17]. The set consists of four *complexity groups* with 0, 1, 3 or 5 unique geons substituted at some position in the chain.

Similar paperclip objects (corresponding to the low complexity set) were used in human psychophysics experiments [2]. In these experiments, subjects were trained with motion sequences of 2-D views, giving an impression of a 3-D object through the kinetic depth effect. In a two-alternative forced choice task on their object set, with single static views of a target or distractor, the miss rate (failure to indicate a match when the target was shown) averaged 30%, indicating that the task is rather difficult. We exceed this rate, but higher complexity objects are more easily distinguished.

Fig. 1. Two silhouette views of the same paperclip object, illustrating the extreme nature of the distortion due to rotation in depth. The left view is the 0°, the right view is

[1] One object in the last group was duplicated in the original set, hence 39 rather than an even 40.

+90°. The raw paperclip image set was provided courtesy of Michael Tarr, Dept. of Psychology, Brown University.

The specific object set used here was also used in a study by Tarr and colleagues attempting to discriminate between view-based and structural theories [4]. In this study no training period was provided; subjects simply had to judge whether two views shown briefly (200 and 100 ms, separated by a mask stimulus) were the same or different. Under these conditions, the baseline set of shapes (all tubes with no geons inserted) were essentially not recognizable by subjects when presented in widely separated views.

4 Results

After a small set of trials to establish the objective function term weights, a set of learning and recognition trials were performed. The number of training views was varied from two to seven, beginning with {-90,-60} views and adding successive views in order. Two variants of the cross entropy approach were performed; first a reference view was chosen, and the partition cell mean occupancies of previous signatures was compared to that view. In the second, the mean partition cell occupancy of all views was compared to the mean of previous signatures. In the single set of experiments, no substantial difference in performance was found for this variation. These two variations were each tested with object worlds consisting of 20 and 39 objects to assess recognition performance scaling with number of objects. During recognition trials, all views of every object were presented to compute error rates; the following figure summarizes the results.

.

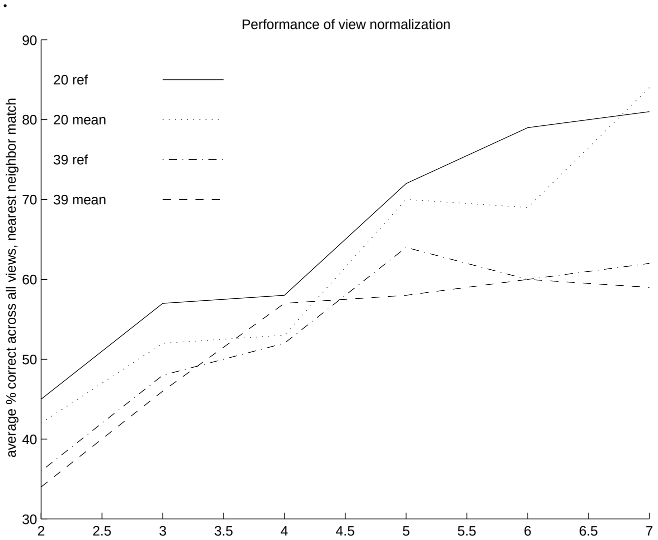

Fig. 2. Summary of nearest neighbor classifier recognition rates with training on 2-7 views, for 20 vs. 39 object worlds, and for reference and mean cross-entropy computations.

We observe that while each of the two stages over time exhibits a general trend toward increasing or decreasing phase space volume, the volume varies non-monotonically from step to step in the transient regime. We tested the performance of a CML with stationary dynamics (i.e. a single stage network) and found it was surprisingly effective, when given the same upper bound of 16 iterations as the two stage network. The following table shows results of 10 trials using the same learning parameters; additional testing is in progress to assess whether these trends are statistically significant. It appears that the underlying normalizing recognizer mechanism can function with a single stage, but that both the best recognition rates and the mean number of iterations (recognition time) are slightly improved by the two stage process.

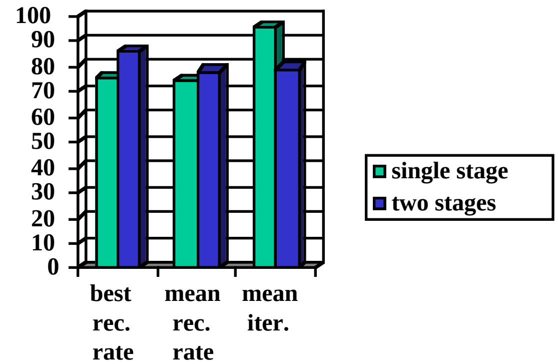

Fig. 3. Best and mean recognition rates and mean iterations as a percent of total allowed iterations for single stage (stationary and two stage (i.e. Soca network) with all seven views provided, 10 trials each with identical learning parameters.

5 Conclusion

The Soca network approach shows promise as a machine vision technique. When adjusted for an equal number of training views the network achieves recognition performance comparable to other recent view based approaches, although to date a common test suite for rotation in depth has not emerged, and the complexity of objects and amount of raw information in images varies [11]. In a single computation specified by few parameters, the network performs both local feature integration (by the diffusion process) and solves the normalization of views problem. In contrast to other approaches, no global statistics of the object world are used to adjust weights or select optimal center views.

The system captures aspects of human psychophysical performance. Both the Soca system and the Chorus RBF ensemble show a decrease in performance with increasing distance from training views when a subset of views is used for training. However, the Soca network shows *variation in the time to reach the population code*. This parameter is implicit in taking iteration counts as a parameter during learning, suggesting that optimal readout of a computation occurs at a particular time or slow wave phase, since the system is not reaching an attractor.

The resulting changes in occupancy of dynamical stages are in accord with observed stimulus related fluctuations in correlation of multi-neuron experiments [18], with the time course of changes in local field potential coherence between primary visual cortex and IT [19], and with stimulus-linked aperiodic oscillations [20]; in addition, we believe observations of stimulus linked slow (4-7 Hz) oscillations in the temporal pole [21] support the present theory; we would predict that slow wave bifurcation and coupling controls would be fed into regions performing the normalization task both in object memory formation and in primed search. Area TEO is the most likely candidate due to direct temporal pole connections, though area V4 has more localized receptive fields like the present system.

The small advantages in reaction time and error rates (Fig. 3) seen in our preliminary comparison of stationary and two stage "Soca-style" dynamics suggest that evolution exploits slow wave bifurcation and coupling control to improve the recognition performance of stationary systems. Further testing is required to ascertain whether these trends are statistically significant; however, more optimization of the learning methods (mutation rates, possible use of Gaussian mutation) should be performed first, so that the variance in recognition rates between trials is minimized prior to such a comparison.

As a biological theory, we bear the burden of explaining observations of specific neurons whose activity increases for specific stimulus in IT regions (for review, see [22], [23]). Two explanations are envisioned. First, such neurons may simply be feeding into a more complex computation which the organism actually uses; the experimenter is able to distinguish the stimulus from a limited range of inputs, but might not be able to do so with a broad range of stimulus. Alternatively, these observations could be indicative of readout assemblies which are sensitive to particular combinations of instantaneous (or short time averaged) oscillation frequencies corresponding to the computation and coding process of a Soca-like network attempting to map different views or other transformations to a common subspace. Observations of multiple active areas with optical recording [24] have been interpreted as activation of RBF style prototype units, situating an object in a representation space spanned by multiple prototypes [25]. We propose that these are multiple readout assemblies, indicating the exploitation of multiple dynamical recognizers to improve the performance, or simply assemblies which correspond to populations close in the representation space by virtue of solving the stimulus identity problem for objects with similar low order statistics. Elucidation of the low level dynamics of interactions between Soca type networks and hypothetical readout assemblies remains an area for future work.

References

[1] S. Duvdevani-Bar and S. Edelman, "Visual recognition and categorization on the basis of similarities to multiple class prototypes," *Intl. J. of Computer Vision*, vol. 33, 1999., pp. 201-228.

[2] H. H. Bulthoff and S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," *Proc. Natl. Acad. Sci. USA*, vol. 89, 1992., pp. 60-64.

[3] M. J. Tarr and S. Pinker, "Mental rotation and orientation-dependence in shape recognition," *Cognitive Psychology*, vol. 21, 1989., pp. 233-282.

[4] M. J. Tarr, H. H. Bulthoff, M. Zabinski, and V. Blanz, "To what extent do unique parts influence recognition across changes in viewpoint?," *Psychological Science*, vol. 8, 1997., pp. 282-289.

[5] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable Multi-Scale Transforms or, "What's Wrong with Orthonormal Wavelets".," *IEEE Trans. Information Theory*, vol. 38, 1992., pp. 587-607.

[6] H. J. Wolfson and L. Yehezkel, "Transformation Invariant Indexing," in *Geometric Invariance in Computer Vision*, J. L. Mundy and A. Zisserman, Eds. MIT Press, Cambridge, 1992.

[7] B. W. Mel, "SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition," *Neural Computation*, vol. 9, 1997., pp. 777-804.

[8] S. Edelman, *Representation and Recognition in Vision*. MIT Press, Cambridge MA, 1999.

[9] W. P. Softky and C. Koch, "The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs.," *Journal of Neuroscience*, vol. 7, 1994., pp. 177-191.

[10] H. Fuji, J. Ito, A. K., I. N., and Tsukada M.. Neural Networks, 1303-1350, 1996, "Dynamical Cell Assembly Hypothesis: Theoretical Possibility of Spatio-temporal Coding in the Cortex.," *Neural Networks*, vol. 9, 1996., pp. 1303-1350.

[11] D. DeMaris, "Synchronization Opponent Networks: Dynamics, Computation and Coding for Similarity and Object Recognition," Dissertation, Electrical and Computer Engineering. University of Texas at Austin, 2001.

[12] L. F. Abbot and T. J. Sejnowski, "Introduction to Neural Codes and Distributed Representations," in *Neural Codes and Distributed Representations: Foundations of Neural Computation*, L. F. Abbot and T. J. Sejnowski, Eds. MIT Press, Cambridge, 1999, pp. vii-xxiii.

[13] A. Del Bimbo, *Visual Information Retrieval*. Morgan Kaufmann, San Francisco, 1999.

[14] K. Kaneko, "Overview of coupled map lattices," *Chaos*, vol. 2, 1993., pp. 279-282.

[15] J. B. Pollack, "The Induction of Dynamical Recognizers," *Machine Learning*, vol. 7, 1991., pp. 227-252.

[16] A. Rosenfeld, *Picture Languages: Formal Models for Pattern Recognition*. Academic Press, 1979.

[17] I. Biedermann, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, 1987., pp. 115-147.

[18] A. M. H. J. Aertsen and G. Gerstein, "Dynamic aspects of neuronal cooperativity: fast stimulus-locked modulations of effective connectivity.," in *Neuronal Cooperativity*, J. Kruger, Ed. Springer-Verlag, Berlin, 1991.

[19] S. L. Bressler, "Large-scale cortical networks and cognition," *Brain Research Reviews*, vol. 20, 1995., pp. 288-304.

[20] B. Richmond, L. Optican, M. Podell, and H. Spitzer, "Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex," *Journal of Neurophysiology*, vol. 57, 1987., pp. 132-146.

[21] K. Nakamura, A. Mikami, and K. Kubota, "Oscillatory neuronal activity related to visual short term memory in monkey temporal pole," *Neuroreport*, vol. 3, 1992., pp. 117-120.

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

[23] K. Nakamura and K. Kubota, "The primate temporal pole: its putative role in object recognition and memory," *Behavioral Brain Research*, vol. 77, 1996., pp. 53-77.

[24] K. Tanaka, "Neuronal mechanisms of object recognition," *Science*, vol. 261, 1993., pp. 685-688.

[25] T. Sugihara, S. Edelman, and K. Tanaka, "Representation of object similarity among three dimensional shapes in the monkey," *Biological Cybernetics*, vol. 78, 1998., pp. 1-7.