Circular Backpropagation Networks Embed Vector Quantization

Sandro Ridella, Stefano Rovetta, and Rodolfo Zunino

***Abstract*—This letter proves the equivalence between vector quantization (VQ) classifiers and circular backpropagation (CBP) networks. The calibrated prototypes for a VQ schema can be plugged in a CBP feedforward structure having the same number of hidden neurons and featuring the same mapping. The letter describes how to exploit such equivalence by using VQ prototypes to perform a meaningful initialization for BP optimization. The approach effectiveness was tested considering a real classification problem (NIST handwritten digits).**

***Index Terms*—Feedforward neural networks, optical character recognition, vector quantization.**

I. INTRODUCTION

Adding a quadratic term to the activation function of linear neurons can greatly enhance the representation ability of multilayer perceptrons (MLP's) without inflating their VC dimension [1]. Circular backpropagation (CBP) networks support both surface-based and prototype-based representations in classification problems; it has been shown that CBP is a unifying model for MLP's and radial basis function (RBF) networks [1]. This letter proves that CBP encompasses vector quantization (VQ) paradigms as well. Thus one can plug VQ prototypes in a CBP network, with the same number of neurons and supporting the same mapping. The fact that CBP structures can repeat the winner-takes-all (WTA) behavior enables one to switch safely from one representation to the other, while preserving a network's mapping function. This property is exploited here to initialize BP training.

The general problem can be stated as follows: a set X of input samples are drawn from a d-dimensional space and belong to one of N_c classes

$$\begin{aligned}&C = \{C_1, \cdots, C_{Nc}\}:\\ &X = \{(\boldsymbol{x}^{(l)}, c^{(l)}), l = 1, \cdots, N_p, \boldsymbol{x}^{(l)} \in R^d, c^{(l)} \in C\}.\end{aligned}$$

A mapping network $T^{(W)}(\boldsymbol{x})$ is instantiated by a set of parameters W which includes elements from both scalar and functional spaces.

A CBP network is a nonrecursive MLP with three layers of neurons (the input, hidden, and output layers). The hidden layer includes N_h nonlinear units; the output one has N_c units (one per class) that are made mutually exclusive by a WTA or Soft-Max operation. The CBP model augments the basic MLP by an additional input, computed as the sum of the squares of the other input values. A CBP network's mapping can be expressed [1] as

$$\begin{aligned}T_{\mathrm{CBP}}^{(f,W)}(\boldsymbol{x}) &= \max_{k=1,\cdots,N_c}\{o_k(\boldsymbol{x}, f, W)\};\\ o_k(\boldsymbol{x}, W) &= v_{k0} + \sum_{j=1}^{N_b} v_{kj} f(-g_j \cdot \|\boldsymbol{x} - \boldsymbol{w}_j\|^2 + \phi_j)\end{aligned} \tag{1}$$

where $f(\ )$ is a nonlinear function—for example, σ-CBP networks use a sigmoidal activation, whereas ρ-CBP networks involve Gaussian RBF's. The weights $\boldsymbol{w}_j, \phi_j$, and g_j set the centroid coordinates, the extent, and the slope of the jth unit's region of influence, respectively.

In the VQ paradigm, each neuron $\boldsymbol{w}_j \in R^d$ is calibrated by a local class information, $\{\alpha_j^{(c)}; c = 1, \cdots, N_c\}$, representing the distribution of class shares for the jth neuron: $\sum_c \alpha_j^{(c)} = 1$. A WTA mapping schema classifies input samples according to a minimum-distance criterion, and associates to each sample the most likely class, $c(\boldsymbol{w}_j) \equiv \max_c\{\alpha_j^{(c)}\}$. The VQ mapping schema is defined as

$$T_{\mathrm{VQ}}^{(W)}(\boldsymbol{x}) = c(\arg\min_{\boldsymbol{w}_j \in W}\{\|\boldsymbol{x} - \boldsymbol{w}_j\|^2\}). \tag{2}$$

Section II proves analytically the equivalence between VQ and two classes of CBP models. Section III describes the application of the overall framework to speed up BP convergence. Conclusions are drawn in Section IV.

II. EQUIVALENCE FRAMEWORK

Stating that CBP encompasses VQ means that, for each set of VQ prototypes, there exists a CBP network whose neurons coincide with VQ units and that classifies each input sample accordingly. To verify this statement, first one builds up a CBP network with as many hidden units as the number of VQ prototypes and with as many output units as the number of classes. The weights in the input-hidden layer are initialized directly with the VQ prototypes' centroids

$$\boldsymbol{w}_j = \boldsymbol{w}_j^{(\mathrm{VQ})}, \quad \phi_j = 0, \qquad \forall j = 1, \cdots, N_h. \tag{3}$$

To initialize the upper layer of weights, set

$$\begin{aligned}v_0 &= 0\\ v_{kj} &= [-1 + 2 \cdot \delta(k, c(\boldsymbol{w}_j))] \cdot \alpha_j^{(k)},\\ &\qquad j = 1, \cdots, N_h,\ k = 1, \cdots, N_c\end{aligned} \tag{4}$$

where $\delta(a, b) = 1$, if $a = b$, and equal to 0, otherwise. The above initialization mirrors VQ calibration: hidden neurons stimulate the output unit associated with the prototype class and inhibit the other ones. The last parameter to be fitted for consistent mapping is the gain g_j of each neuron; for simplicity, a common gain value is assumed for all the neurons: $g_j = g\ \forall j = 1, \cdots, N_h$. Such initializations set up the framework for the equivalence theorems.

Theorem 1 (ρ-CBP's Embed VQ Networks.): Let X be a sample set and $T_{\mathrm{VQ}}^{(W)}(X)$ be the VQ-based mapping schema (2) over X. For each choice of W, there exists a CBP network parametrization, according to (3) and (4), W' such that

$$T_{\mathrm{VQ}}^{(W)}(X) = T_{\mathrm{CBP}}^{(f,W')}(X)$$

with $f(\ )$ = Gaussian RBF.

Proof: The proof of Theorem 1 is constructive. Conditions (3) imply immediately that $W \subset W'$. The RBF activation of each hidden unit can be written as $f_j(\boldsymbol{x}) = \exp(-g \cdot \Delta_j^2(\boldsymbol{x}))$, $j = 1, \cdots, N_h$, where $\Delta_j^2(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{w}_j\|^2$. After initializing the CBP network according to (4) and using the above activation function, the mapping model (1) in the output layer is equivalent to a VQ-based WTA mapping if and only if

$$\begin{aligned}&\alpha^{(c(\boldsymbol{w}_{j^*}))} \exp(-g \cdot \Delta_{j^*}^2(\boldsymbol{x}))\\ &\qquad > 2 \sum_{h \neq j^*} \alpha^{(c(\boldsymbol{w}_{j^*}))} \cdot \exp(-g \cdot \Delta_h^2(\boldsymbol{x})), \qquad \forall \boldsymbol{x} \in X\end{aligned} \tag{5}$$

Manuscript received February 2, 1999; revised April 29, 1999.

The authors are with DIBE–Department of Biophysical and Electronic Engineering, University of Genoa, 16145 Genoa, Italy.

Publisher Item Identifier S 1045-9227(99)05763-X.

1045–9227/99$10.00 © 1999 IEEE

where j^* indicates the best matching neuron, i.e., the unit for which $\Delta_{j^*}^2(\boldsymbol{x}) \leq \Delta_h^2(\boldsymbol{x}),\ \forall h = 1, \cdots, N_h$. Let us now define $\Delta_{II}^2(\boldsymbol{x})$ as

$$\Delta_{II}^2(\boldsymbol{x}) = \min_{h \neq j^*} \{\Delta_h^2(\boldsymbol{x})\}.$$

Considering that $\alpha_h^{(u)} \geq 0$, a consistent bound to condition (5) is given by

$$\exp(-g \cdot \Delta_{j^*}^2(\boldsymbol{x})) > \frac{2(N_h - 1)}{\alpha^{(c(\boldsymbol{w}_{j^*}))}} \exp(-g \cdot \Delta_{II}^2(\boldsymbol{x})), \qquad \forall \boldsymbol{x} \in X \tag{6}$$

Solving (6) gives the gain value that ensures a correct WTA mapping in the output layer for sample $\boldsymbol{x}$. Finally, the correct mapping for all the samples in X can be attained by choosing

$$g = \max_{\boldsymbol{x} \in X} \left\{ \frac{\ln[2(N_h - 1)] - \ln(\alpha^{(c(\boldsymbol{w}_{j^*}))})}{\Delta_{II}^2(\boldsymbol{x} - \Delta_{j^*}^2(\boldsymbol{x})} \right\}. \tag{7}$$

Expression (7) demonstrates that there exists a ρ-CBP network (using only VQ centroids) that performs the same classification of all the samples, hence the equivalence of the two mappings is proved.Q.E.D.

Theorem 2 (σ-CBP's Embed VQ Networks): Let X be a sample set, and let $T_{\mathrm{VQ}}^{(W)}(X)$ be a VQ-based mapping schema (2) over X. For each choice of W, there exists a σ-CBP network parametrization, according to (3) and (4), W' such that $T_{\mathrm{VQ}}^{(W)}(X) = T_{\mathrm{CBP}}^{(f, W')}(X)$, with $f(\ ) =$ sigmoidal function.

Proof: By using the same conventions as above, a neuron's sigmoidal activation function is expressed as $f_j(\boldsymbol{x}) = [1 + \exp(-g \cdot \Delta_j^2(\boldsymbol{x}))]^{-1},\ j = 1, \cdots, N_h$. The sigmoid supports a consistent mapping of distances when $g < 0$, hence the notation $|g|$ will be used in the function's argument for simplicity. A correct WTA behavior in the output layer is attained when an input sample $\boldsymbol{x}$ activates the proper hidden unit in such a way that

$$\begin{aligned} &\alpha^{(c(\boldsymbol{w}_{j^*}))} \cdot [1 + \exp(|g| \cdot \Delta_{j^*}^2(\boldsymbol{x}))]^{-1} \\ &\quad > 2 \sum_{h \neq j^*} \alpha^{(c(\boldsymbol{w}_{J^*}))} \cdot [1 + \exp(|g| \cdot \Delta_h^2(\boldsymbol{x})]^{-1}, \qquad \forall \boldsymbol{x} \in X \end{aligned} \tag{8}$$

Let us now define

$$\Delta_{II}^2(\boldsymbol{x}) = \min_{h \neq j^*} \{\Delta_h^2(\boldsymbol{x})\}.$$

Similarly to Theorem 1, one bounds the right term in condition (8) accordingly and uses the properties $\exp(|g| \cdot \Delta^2) > 1$ and $\alpha_h^{(u)} \geq 0$. Simple transformations give the eventual gain value

$$|g| = \max_{\boldsymbol{x} \in X} \left\{ \frac{\ln[4(N_h - 1)] - \ln(\alpha^{(c(\boldsymbol{w}_{j^*}))})}{\Delta_{II}^2(\boldsymbol{x}) - \Delta_{j^*}^2(\boldsymbol{x})} \right\} \tag{9}$$

which guarantees a correct sample mapping and completes the proof. The equivalence property of ρ-CBP networks holds for σ-CBP ones: this is not surprising, as the latter have been proved to be a superset of RBF networks [1]. Q.E.D.

Theorems 1 and 2 prove that CBP supports VQ mapping on a finite sample set X; one might wonder whether the equivalence holds for an arbitrarily large cardinality of the sample set. Consider a "critical" point $\boldsymbol{x}$ lying at distance ε from the boundary of the space partition pertaining to a VQ prototype. When applying Theorems 1 and 2, the limit condition $\varepsilon \to 0$ implies $|g| \to \infty$ in (7) or (9). Thus a finite gain is associated with a distance threshold, marking a "neutral" stripe running along partition boundaries. When dealing with large sample sets, one first imposes a tolerance on the number of "undecided" samples; then one determines the smallest distance ε^* that satifies the constraint, and designs the CBP gain accordingly using the above theorems.

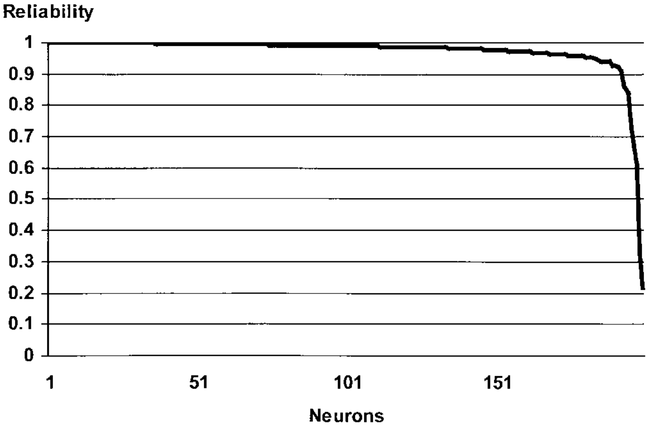


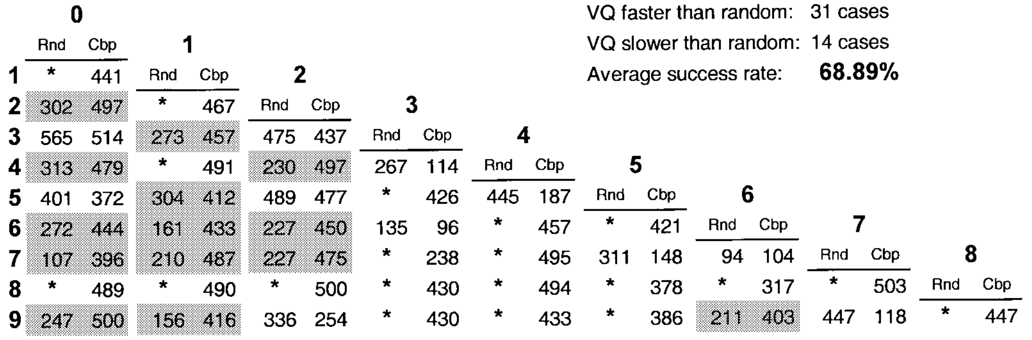
Fig. 1. NIST digit testbed: calibration results for validating VQ-based initialization.

III. Practical Exploitation of the CBP-VQ Mapping Equivalence

A. Practical Network Initialization

In principle, the mapping equivalence between CBP and VQ can operate in two ways. In other words, first one may perform BP training, then one may use weight-reversal expressions [1] to inspect the positions of VQ prototypes. This approach, however, requires careful interpretations of the final gain values and of the interactions among prototypes. Conversely, a possibly easier exploitation of the equivalence is to let VQ neurons initialize the weights of a CBP network. This process is theoretically admitted by the theorems proved in the previous section, and can be justified by pattern-recognition purposes. In VQ classification (2), each Voronoi region V_j has a prototype $\boldsymbol{w}_j$ and is labeled by the predominant class among the samples contained in the region itself. WTA-based class assignment does not depend on the specific position of a sample in V_j; thus VQ calibration can be regarded as a uniform approximation at the local level for the class probability. Supervised VQ-training algorithms (e.g., the LVQ [4] family) can be adopted to best fit the underlying classification task. Otherwise, first one may follow an unsupervised strategy to approximate the overall sample distribution [2]–[5], and then one may calibrate VQ partitions using class information. In fact, the observation of class shares $\{\alpha_j^{(c)}\}$ in each region V_j can give some hint about how homogeneous class distributions are within the VQ-derived partitions. Calibration proceeds locally at the partition level, and does not take into account neighboring partitions. In principle, unsupervised training does not allow any prediction about the classification performance. In fact, a prediction of the classification performance is possible if one can detect "peaks" in the sample density ("clusters") and use the "valleys," separating these peaks, for defining the cluster boundaries [6]. Nevertheless, the analysis of local estimates $\{\alpha_j^{(c)}\}$ can give the opportunity to inspect the distribution of classes within each region. This provides a useful tool for assessing the overall quality of the initialization process; therefore, one might entirely reject VQ-based initialization should local approximations prove unsatisfactory (e.g., if the overall classification error resulting from VQ is too large). The initialization procedure applies independently of the specific VQ-training algorithm, and can be outlined as follows.

1) *Train* a set of N_h prototypes by using a VQ algorithm.
2) *Calibrate* VQ prototypes by evaluating the class distributions on the training set: $\{\alpha_j^{(c)}\}, j = 1, \cdots, N_h$.



	0		1		2		3		4		5		6		7		8	
	Rnd	Cbp	Rnd	Cbp	Rnd	Cbp	Rnd	Cbp	Rnd	Cbp	Rnd	Cbp	Rnd	Cbp	Rnd	Cbp	Rnd	Cbp
1	*	441																
2	302	497	*	467														
3	565	514	273	457	475	437												
4	313	479	*	491	230	497	267	114										
5	401	372	304	412	489	477	*	426	445	187								
6	272	444	161	433	227	450	135	96	*	457	*	421						
7	107	396	210	487	227	475	*	238	*	495	311	148	94	104				
8	*	489	*	490	*	500	*	430	*	494	*	378	*	317	*	503		
9	247	500	156	416	336	254	*	430	*	433	*	386	211	403	447	118	*	447

VQ faster than random: 31 cases
VQ slower than random: 14 cases
Average success rate: **68.89%**

Fig. 2. NIST digit testbed: convergence of VQ-based and random initializations for all class pairs.

TABLE I
VQ CALIBRATION RESULTS

Class	Number of Prototypes
0	15
1	9
2	26
3	19
4	25
5	25
6	17
7	17
8	28
9	19

3) *If* the calibration result is not satisfactory,
 Reject VQ-based weights and *Abort* the initialization.
4) *Build* a feedforward CBP network by using initializations (3) and (4).

Such initialization methodology raises several issues. First, the theorems proved in Section II guarantee that the error rate resulting from BP training will not exceed that obtained by VQ calibration (as the initial classification errors of the two models coincide). If such a rate is low, in practice this property increases the probability of placing the BP starting point in a "good" basin. In this respect, it is worth recalling that the theorems follow a worst case analysis and, for example, do not imply the possibility that several hidden units may contribute to a correct classification. Therefore, in practice the proposed initialization proves too strict, and the BP algorithm implicitly relaxes the WTA constraint by letting hidden units cooperate.

Another crucial issue concerns the shapes of the neurons' activation regions. CBP supports hyperspherical surfaces, hence the result of BP training also includes radii and gains; the latter express boundary sharpness, and the former convey the extent of a neuron's spherical region. By contrast, the shapes of VQ-derived Voronoi regions are arbitrary, hence CBP may not seem the most effective model to exploit VQ-based initialization. This issue can be taken into account in various ways. For instance, one may use a VQ-training algorithm that intrinsically leads to hyperspherical regions (e.g., the method described in [2] and [6]), thus making the equivalence with CBP also hold from a topological perspective. Conversely, the representation ability of the circular model can be augmented by additional second-order terms, yielding hyperelliptical boundaries [7]. The enhancement would best fit the natural convexity of Voronoi regions; on the other hand, the more complex solution might compromise the model's limited VC-dim.

The present research adopted the Plastic Neural Gas algorithm [5] at Step 1. This VQ method, which minimizes the mean-square error over training samples and leads to classical Voronoi structures, has been chosen because it estimates both the proper number N_h of prototypes and their positions at the same time. As a result, in this case, the CBP-VQ equivalence also gives indirectly a hint about the number of hidden neurons in the feedforward network. In the simple generalized-XOR testbed, experimental evidence indicates that, on average, using VQ prototypes speeds up the convergence of CBP optimization by one order of magnitude.

B. Real Domain Test: The NIST Digit Database

In the case of handwritten digits drawn from the NIST database, the original pictures, after normalization and orientation, were mapped into a 140-dimensional feature space. Such a feature-extraction process was obtained through the courtesy of Elsag Bailey SpA [8]. Thus the 60 000 training samples, belonging to $\mathcal{R}^{140}$ were first processed by the Plastic VQ algorithm [4], which yielded $N_h = 200$ prototypes; their class distribution after calibration is given in Table I. The graph in Fig. 1 shows the sorted values of the "reliability," α_j, of each prototype's label, defined as $\alpha_j = \max_c\{\alpha_j^{(c)}\}$. Most of neurons exhibit singular local distributions $\alpha_j \geq 0.8$; the overall classification error resulting from VQ (about 1.79%) seems quite interesting, considering the unsupervised training and the multiclass problem nature. This result gave the operational basis for applying the VQ-based initialization.

The convergence rate and speed provided by random initialization were so low that a direct comparison with the VQ-based initialization performance is unfeasible; in fact, the random networks never succeeded in attaining a smaller classification error than that associated with the CBP-VQ MLP before 10 000 epochs. A quantitative evaluation of the initialization method was obtained by a set of simpler tests, also in view of the huge computational effort involved. For each possible pair of classes, the related samples were extracted from the database and formed a limited training set; the results yielded by the VQ-based initialization (Table I) were compared with those obtained by a set of ten random networks trained on the same data subsets. Training runs stopped when attaining correct classification of all samples, or when reaching a limit on the number of epochs. The algorithm used for BP optimization (AMBP) is presented in [9].

Fig. 2 gives the best case number of epochs at convergence for the random networks and the corresponding performance of the VQ-initialized network. The $*$ marks indicate either failure or convergence beyond 1000 epochs; the grayed cells point out the unsuccessful cases in which random initialization prevailed. VQ-based initialization performed better than the best random case in about 69% of cases. This represents a satisfactory result also given the complexity of the problem involved. In practice, the proposed initialization allows one

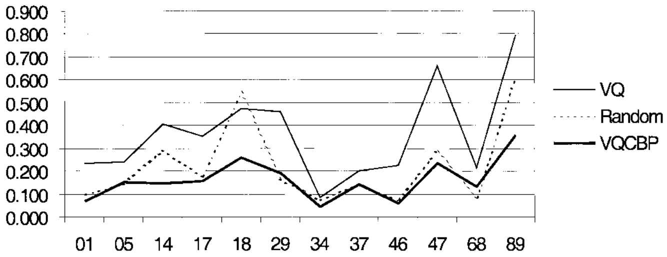


Fig. 3. Comparison of generalization errors (y axis) for various pairs of digits (x axis).

to refine the basic VQ training process even to a zero training error by a standard BP optimization at a limited cost.

A crucial issue about the practical relevance of the results concerns the generalization performance of the trained networks. In fact, the VQ-based initialization does not affect the classifier's VC-dim, hence theory does not predict a difference in generalization ability with respect to the randomly initialized networks. On the other hand, as the VQ classifier *per se* delivered a rather high expressive power (>98%), one might concern that further training just stimulated overfitting phenomena. Therefore, a different "test" set of 60 000 handwritten digits was used to compare the overall classification errors for the various initializations empirically. The comparison considered the performances of the VQ (alone), of the VQ + CBP, and of the randomly initialized network, respectively.

Fig. 3 presents graphically a sample of the generalization results; similar achievements were observed for all possible pairs of digits and are not reported for brevity. Empirical evidence pointed out a significant reduction in generalization error when enhancing the basic VQ classifier by means of CBP training; the increase in performance appears quite satisfactory when considering the application domain, in which enhancing overall generalization accuracy beyond 97% often proves very difficult. Finally, the entire digit test set (10-class problem) was processed by the huge networks including the 200 prototypes and involving all training samples; the measured generalization errors (VQ alone: 1.83%, random-init CBP: 0.79%, VQ-init CBP: 0.785%) confirm the results obtained in the dual-class subproblems.

Thus experimental data validate the proposed initialization method, as its ultimate effect is to speed up convergence without affecting generalization ability. In order to explain intuitively such a result, we conjecture that VQ-based initialization is not merely effective in decreasing the initial training error, but also provides the optimization process with a "reasonable" starting point that ultimately enhances generalization performance.

IV. CONCLUSIONS

The unifying view of the MLP and VQ fields opens new and interesting vistas for integrated neural models, in particular, for training algorithms. This letter has described an analytical technique to initialize MLP weights with VQ prototypes; the method's validity was confirmed experimentally in a complex domain. Clearly, as is the case with any initialization technique, it cannot be guaranteed that the proposed procedure will apply to any classification problem. Nevertheless, a specific advantage of the methodology lies in the possibility to evaluate the quality of a particular initialization phase in advance and possibly to reject it altogether. It is worth noting, however, that in practice the method operates successfully in most domains featuring "reasonable" sample distributions.

REFERENCES

[1] S. Ridella, S. Rovetta, and R. Zunino, "Circular back-propagation networks for classification," *IEEE Trans. Neural Networks*, vol. 8, no. 1, pp. 84–97.
[2] M. M. Van Hulle, "Kernel-based equiprobabilistic topographic map formation," *Neural Comput.*, vol. 10, pp. 1847–1871, 1998.
[3] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.
[4] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. New York: Springer Verlag, 1989.
[5] S. Ridella, S. Rovetta, and R. Zunino, "Plastic algorithm for adaptive vector quantization," *Neural Comput. Applic.*, vol. 7, no. 1, pp. 37–51, 1998.
[6] M. M. Van Hulle, "Density-based clustering with topographic maps," *IEEE Trans. Neural Networks*, vol. 10, no. 1, pp. 204–206, Jan. 1999.
[7] D. P. Casasent, "Multifunctional hybrid neural net," *Neural Networks*, vol. 5, no. 3, pp. 361–370, 1992.
[8] A. M. Colla and P. Pedrazzi, "Binary digital image feature extracting process," U.S. Patent 5 737 444, Apr. 7, 1998, Assignee: Italian Ministry for Univ. and Sci. Res.
[9] G. P. Drago, M. Morando, and S. Ridella, "An adaptive momentum back propagation (AMBP)," *Neural Comput. Applic.*, vol. 3, pp. 213–221, 1995.

New Stability Conditions for Hopfield Networks in Partial Simultaneous Update Mode

Donq-Liang Lee

***Abstract*—Cernuschi-Frías has proposed a partial simultaneous updating (PSU) mode for Hopfield networks. He also derived sufficient conditions to ensure global stability. In this letter, a counter-example is given to illustrate that the PSU sequence may converge to limited cycles even if one uses a connection matrix satisfying the Cernuschi-Frías conditions. Then, new sufficient conditions ensuring global convergence of a Hopfield network in PSU mode are derived. Compared with the result of fully parallel mode case, the new result permits a little relaxation on the lower bound of the main diagonal elements of the connection matrix.**

***Index Terms*—Global stability, Hopfield network.**

I. INTRODUCTION

The Hopfield network [1], [2] is one of the famous neural networks with a wide range of applications, such as content addressable memory [2], pattern recognition [1], and combinatorial optimization [10]. In the synthesis of such a network, ensuring a convergence of the state trajectories starting from arbitrary initial state to a fixed point is of particular importance. Such a convergence property is the basis for the potential applications of the network. Afterwards many researchers have focused on the following two distinct update modes: 1) asynchronous (or serial) mode, in which a neuron is chosen

Manuscript received January 26, 1999. This work was supported by the National Science Council, Taiwan, R.O.C., under Grant NSC88-2213-E-233-001.

The author is with the Department of Electronic Engineering, Ta-Hwa Institute of Technology, Chung-Lin, Hsin Chu, Taiwan 307, R.O.C.

Publisher Item Identifier S 1045-9227(99)06052-X.