Harold Szu, Ph.D.

Naval Research
Laboratory
Washington, D.C.

Dr. Szu received his Ph.D. from Rockefeller University in 1971, and from 1972 to 1974, he taught at the University of North Carolina at Chapel Hill. He served 5 years in the Plasma Physics Division of NRL, and obtained two patents for (1) an infrared free-electron laser, and (2) a superconducting heavy-ion accelerator. In 5 years at the Optical Sciences Division of NRL, Dr. Szu obtained two patents for (1) a phase-conjugate-crystal interferometer, and (2) a computer-generated-hologram pattern classifier. Two patents (pending) are a result of 3 years in Tactical Electronic Warfare at NRL, (1) a nonconvex optimization Cauchy machine, and (2) superconducting sensor arrays, neuromorphic computers, and triode amplifiers. In addition to more than 100 technical papers and reports, Dr. Szu has published three books: "Fluctuation-Dissipation Theorems," "Optical and Hybrid Computing," and "Pattern Recognition Based on Human Visual Systems and Neural Networks" for Lawrence-Erlarbaum. Dr. Szu also teaches Adaptive Neural Networks at UCLA, serves as a governing board member and officer of the International Neural Network Society (INNS), and is the editor of Pergamon Journal, Neural Networks.

## WHAT IS THE SIGNIFICANCE OF NEURAL NETWORKS FOR AI?

### Abstract

Associative memory (AM) and attentive associative memory (AAM) have been reviewed in terms of simple neural networks (both uniform and nonuniform matched filter banks - read by inner products and written by outer products in parallel). Whereas AM has been applied to optical character recognition (OCR) using the set of orthogonal feature vectors deduced from image processing and computer vision, AAM can incorporate AI expert system techniques for determining the nonuniform linear combination of outer products. A rule-based system can more efficiently incorporate the frequency distribution of distorted characters according to user group profiles; i.e., left-handed versus right-handed writing. Specifically, in this paper we have examined the degree of fault tolerance in AM, the ability of generalization by interpolation (auto-associative memory), and abstraction by extrapolation (hetero-associative memory). The efficiency of the closed system of rule-based knowledge representation of AI using tuple storage has been combined with the flexibility of the non-rule-based open system using the matrix knowledge representation of NI (coined for either neural, or network, or natural intelligence). Thus, the ability of generalization and abstraction becomes possible in a combined intelligent system of AI and NI.

# WHAT IS THE SIGNIFICANCE OF NEURAL NETWORKS FOR AI ?

by Harold H. Szu
Naval Research Laboratory, Code 5756
Washington D.C. 20375

ABSTRACT

Associative memory (AM) and attentive associative memory (AAM) have been reviewed in terms of simple neural networks (both uniform and non-uniform matched filter banks: read by inner products and write by outer products in parallel). While AM has been applied to the optical character recognition (OCR) using the set of orthogonal feature vectors deduced from image processing and computer vision, AAM can incorporate AI expert system techniques for determining the non-uniform linear combination of outer products. A rule-based system can more efficiently incorporate the frequency distribution of distorted characters according to user group profiles, say left-handed writing versus right-handed writing. Specifically in this paper, we have examined the degree of fault tolerance in AM, the ability of generalization by interpolation (auto-associative memory) and abstraction by extrapolation (hetero-associative memory). The efficiency of the closed system of rule-based knowledge representation of AI using the tuple storage has been combined with the flexibility of the non-rule based open system using the matrix knowledge representation of NI (coined for either Neural, or Network, or Natural Intelligence). Thus, the ability of generalization and abstraction becomes possible in a combined intelligent system of AI and NI.

## 1. INTRODUCTION

The question of the significance of neural networks for AI may be subdivided into three aspects.

(i) How can neural networks help solve AI problems ?

ANSWER: Both the well understood fault-tolerance of associative memory (AM), and the lesser understood ability of neural networks for generalization and abstraction, can be usefully incorporated into AI techniques.

(ii) How can AI help solve neural network problems ?

ANSWER: Similar to computer aided design, AI expert systems with a neural network modules can help design special purpose architectures for neural network computing.

(iii) What unsolved problems can be solved efficiently by combining AI and NI (coined for either Neural, or Network, or Natural Intelligence) techniques to utilize their respective strengths?

ANSWER: The optical character recogniton (OCR) for reading hand-written bank check and zip-codes, can be solved by combining both AI and NI techniques, as described in this paper.

Because we can only build a small neural network, we wish to endow a small se neurons with a human-like intelligence. With present technology, whether it be electronic optical, one cannot build a neural network of more than several hundred neurons, using existi processor elements (PE's), because of the technological difficulty associated with der interconnectivity, about $N^2$ for N PE's. Thus, artificial neural networks can not yet match the si and the complexity of the human brain, that has billions of neurons and thousands interconnects for each neuron. If we are *not*, overly ambitious in developing a **general purpo** neural computer, we can built a **special purpose** neural computer for solving special purpo problems, such as OCR.

One way to accomplish this special purpose neural computer is to combine the traditior rule-based AI wisdom with non-rule-based NI learning. This is particularly desirable in solvi OCR problems because the available small neural networks can use better feature vectors obtain from other disciplines. Neural networks, built with current technology, can then provide fa tolerance for input feature vectors variations. The specific problem of hand-written charact recognition, differs from the more regular, hand-printed, alphanumeric recognition problem that it must account for such complications as connected characters and characters broken segmentation.

Conceptually, one could solve the OCR problem using analytic, rule-based AI or neu network techniques. The OCR problem can be subdivided into character (or character strin statistics, font recognition, and character recognition; the most efficient techniques for these thr subproblems are analytic (statistical), rule-based AI, and neural networks, respectively. Since t statistical techniques, applied to alphanumeric frequencies, is well known, this topic will not discussed further. In solving the font recognition subproblem, AI rules can be set by the (statisti frequency distribution of individual distorted characters according to user group profiles, e.g. le handed writing versus right-handed writing. It is efficient to design AI expert system that dra upon the classical statistical pattern recognition, e.g. one stroke difference exists between "*P* " a "*R* ", or between"*O* " and "*Q* ", or in a low pass filter viewpoint only one stroke location difference exists among four rounded letters "*P* " and "*R* ","*O* ", and "*Q* ". Furthermore, the rules of **pair character distortion distribution** can help solve the problem of ***connected charac*** and ***broken character after segmentation.*** such as **two scripted zeros.** The **pair characer correlati matrix can be analyzed by the technique of the Karhunen-Loeve** procedure in image processir The Karhunen-Loeve technique is compatable with AM's outer product decomposition. With t help of AI rule-based system, both the first and the second order statistics can be incorporated in t formalism of **attentive associative memory** (AAM), that processess the extra degrees of freedom the non-uniform storage of vector outer products based on a given set of critical feature vectors.

Because the open-ended knowledge of input pattern variations may be efficien controlled by using other disciplinary knowledge, such as AI and computer vision with a result better combined technology, we shall review AM and AAM, and various OCR approaches means of their specific techniques used for feature extraction and techniques used for gro classification. The sooner we accept implementation limitations of the present neurocompute the better we can work with other disciplinary researchers. For example, we can work wi researchers in AI, computer vision, image processing. Since this cross disciplinary collaboration

by nature not easy because of different trainings and languages involved, then this paper may serve a door opener for both.

Pattern recognition reseachers have been successful in machine-printed character recognition (CR) compared to optical character recognition (OCR) of hand-written bank checks or zipcodes. Difficulties of applying AI alone to an intelligent OCR may be due to the lack of non-rule-based capability of generalization and abstraction. This may be constrained by the traditional AI **one dimensional (1-D) knowledge representation**, e.g. an ordered set of tuples used in semantic networks. Similarly, difficulties of applying the neural network alone to an intelligent OCR may be in selecting critical features that is precisely one of the most challenging and unsolved problems (others are segmentations and locations). On the other hand, AI is efficient in reduce the problem to a sub-problem based on **1-D** knowledge representation of simple rules, and NI provides the fault-tolerant OCR system based on **2-D** knowlege representation. Together they give the possibility of generalization and abstraction. Thus, Szu and Tan (1988) have considered a less risky approach that consists of the traditional AI researchers who know about OCR critical features, and the neural network experts who know about AM fault tolerance. Technological developments have pointed to the readiness of such collaborations, since 2-D storage by chips or optical disks becomes cheaper than the traditional 1-D content addressable memory (CAD) processor. What's needed is a smart coprocessor such as neurocomputer. As a matter of fact, due to the 2-D nature of light, optical expert systems based on AM have been designed by Szu and Caulfield (1987) who have shown as simple replacement of 1-D tuples by 2-D matrices in a semantic network the alias problem for data fusion is solved by matrix addition and thresholding. The opto-electronical implementation of attentative associative memory model of Athale, Szu & Frielander (1986) can be expanded by means of a priori probability compiled by a pair-character correlation function of script letters. These papers may facilitate both sides the starting line of collaborations.

In this paper, we have reviewed the orthogonal subspaces of features and examined (1) the degree of fault tolerance , (2) the generalization by interpolation to other orthogonal feature vectors within the subspace, and (3) the abstraction by extrapolation to other subspaces. **AAM** may be formulated by a linear combination of outer products based on a set of orthogonal feature vectors. The combination coefficient is called the attention parameter, because it enters into the eigenvalue of **AAM** matrix that governs the recall convergence. We review briefly about the dynamics of attentive associative memory published by Szu (1988) elsewhere using arbitrary coefficients. In this paper we explicitly introduce a AI-tuple for the attention vector $\mathbf{a} = \{a_n , n=1,...M\}$, where the inner product between the difference vector between an averaged stochastic input $|Q>$ and a fixed memory state $|m>$ is naturally used as the attention parameter defined in terms of Dirac's inner product notation: $a_m = <m|m> - <m|Q>$. Such an **AAM matrix** has non-white eigenvalue spectrum $\lambda_n \equiv a_n \cdot (A / B)$ where the attentive memory capacity is $A \equiv \Sigma^M_{n=1} a_n$, and B is the length of the feature vectors (e.g. the number of bits). Iterative recalls are used. **Paying non-uniform attention** ($a_n \geq 1$) **increases the memory capacity** $A \geq M$ **together with a faster convergence rate proportional to the larger eigenvalue** $\lambda_m \geq \lambda$ **than a uniform attention**( i.e.$a_m$ = 1). Szu's (1988) analysis has suggested that the eigenvalue spectrum and its dithering by input ensemble can play a crucial role for the convergence associated with a nonlinear dynamical system.

## 2. Associative Memory

Matrix associative memory works like a parallel bank of matched filters but much m efficiently in at least three counts: (1) no address coding of input and decoding for output necessary , (2) operations are done in parallel, and (3) the connectivity matrix can be determi by itself using various adaptive (learning) algorithms.

An analytical and numerical example of AM is given as follows:

We denote M feature vectors as binary words, $U^{(m)}$, m=1,...M. Each word has B bits. inner product of Eq(1) measures the norm, the number of bits that are one.

$$U^T \bullet U = \#\text{ of one's} \qquad (1)$$

where the superscript transpose the column vector to a row vector.

The associated **bipolar words**, denoted by $V^{(m)}$, m=1, ...M, are defined as follows:

$$V = (2\,U - 1) = Sgn(\,U\,) \qquad (2)$$

where the unit vector **1** has all entries equal 1 and Sgn is the sign function that changes zero negative quantities to -1. We prefer bipolar version to binary version because : (1) the in product norm is always identical to the number of bits, B:

$$V^T \bullet V = B = \langle V | V \rangle \ , \qquad (3)$$

rewritten here in terms of Dirac's bracket notation: <bra|ket> for the inner and |ket><bra| for outer product, (2) the nature of "exclusive or" can be easily represented by bipolar multiplication

$$+1 \times +1 = 1, -1 \times -1 = 1, +1 \times -1 = -1, -1 \times +1 = -1,$$

(3) the inner product norm is related to the Hamming distance, defined to be the number different bits between two vectors no matter where the differences occur.

We assume an orthogonal set of feature vectors defined as follows:

$$V^{(n)T} \bullet V^{(m)} = B\ \delta_{n,m} = \langle n | m \rangle \qquad (4)$$

where $\delta_{n,m}$ is the Kronecker delta. The outer product weight matrix **W** represents au associative memory:

$$[\,W\,] = \Sigma_m\, [\,V^{(m)} V^{(m)}] = \Sigma_m\, |m\rangle\langle m| \qquad (5)$$

Hopfield (1982, 1984) assumed the auto-associative matrix [T] to be traceless. That was used together with the symmetry property to prove convergence. Thus, the second term of Kronecker's delta matrix (1's along the main diagonal and zero elsewhere) is introduced in Eq (6) to make it traceless.

$$B[T]_{ij} = [W]_{ij} - M\,\delta_{i,j} \tag{6}$$

B is the normalization constant, and M is the memory capacity. Using the trace operation denoted by Tr, we can easily verify Eq (6) to be traceless.

$$Tr(\,|m><m|\,) = B \tag{7}$$

$$Tr([\delta_{i,j}]) = B \tag{8}$$

The tradeoff between the memory capacity and the degree of fault-tolerance has been estimated to be about 15 % of B bits [Hopfield (1982)] for pseudo-orthogonal vectors. That is,

$$M = 0.15\,B \tag{9}$$

For orthogonal feature vectors, however, the capacity is 100 %.

$$M = B \tag{10}$$

This fact can be demonstrated by the eigenvalue problem of the matrix which is defined to be

$$[T]\,|n> = \lambda_n\,|n> \tag{11}$$

where the eigenvalue can be easily verified, using Eqs (4) and (6), to be *degenerate* , namely, a white spectrum for all M states,

$$\lambda_n = 1 - (M/B) \tag{12}$$

The full capacity, M = B, corresponds to a zero eigenvalue for all B orthogonal eigenstates, one for each feature vector.

Consider a simple example where B = 4. There are 4 possible orthogonal vectors and $2^4$ = 16 possible words denoted by:

0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15

We introduce orthogonal subspaces defined by the number of contiguous 1's in the binary word. The subspace consisting of words 13, 11, 7, and 14 is obviously orthogonal by shifting a "one" among 3 zeroes from the the left to the right end of the word.

| Word | Binary p | Word comple. | | Binary p | | | Word Bipolar | Word Bipolar comple. | |
|---|---|---|---|---|---|---|---|---|---|
| 13 | 1101 | 2 | 0010 | 3 | 13 | 1 1 -1 1 | 2 | -1-1 +1-1 | 3 |
| 11 | 1 011 | 4 | 0100 | 3 | 11 | 1 -1 1 1 | 4 | -1+1-1 -1 | 3 |
| 7 | 0111 | 8 | 1000 | 3 | 7 | -1 1 1 1 | 8 | +1-1-1-1 | 3 |
| 14 | 1110 | 1 | 0001 | 3 | 14 | 1 1 1 -1 | 1 | -1-1-1+1 | 3 |
| 15 | 1111 | 0 | 0000 | 4 | 15 | 1 1 1 1 | 0 | -1-1 -1 -1 | 4 |
| 6 | 0110 | 9 | 1001 | 2 | 6 | -1 1 1-1 | 9 | +1-1-1+1 | 2 |
| 12 | 1100 | 3 | 0011 | 2 | 12 | 1 1 -1-1 | 3 | -1-1+1+1 | 2 |
| 10 | 1010 | 5 | 0101 | 1 | 10 | 1-1 1-1 | 5 | -1+1-1+1 | 1 |

It is readily verified that the subspace of bipolar words (13, 11, 7, 14) are mutua orthogonal to one another, as shown in Figure 1. They happen to be related to the Walsh functi of periodicity p=3. The corresponding binary words have an equal angle among them [$\cos^{-1}$ (2/ that is not 90°. Also, the second subspace of bipolar words (15, 6, 12, 10) are also orthogonal but two subspaces are *not* orthogonal to each other.

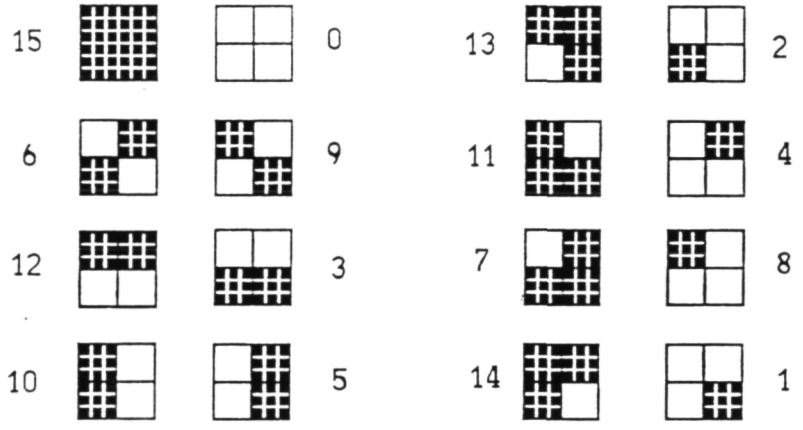



Figure 1. Two-Dimensional Representation of Walsh Base Functions Used to illustrate the fault tolerance and generalization properties of Associative Memory

We consider the storage of one word in memory.

$$4\,[T_1] = [13] = |13\rangle\langle 13| - \delta \qquad (13)$$

If the outer product is properly normalized, it is related to the projection operator:

$$[P] = \delta - |13\rangle\langle 13| \quad (1/B) \tag{14}$$

Using Eq (4) , it can be verified that

$$[P]^2 = [P]. \tag{15}$$

We will show (1) the ability of fault tolerance, and (2) the ability for generalization.

**Fault Tolerance**

The following sequence of erasing (zero out) successively from the bipolar bits illustrate tolerance of missing bits.

(1) one missing bit

$$[13](0\ 1\text{-}1\ 1)^T = Sgn(3\ \text{-}2\ \text{-}2\ 3)^T = |13\rangle \tag{16}$$

where Sgn is sign function representing the sigmoid neuron response by the point nonlinearity extracting the algebra sign of each entries.

(2) two missing bits

$$[13](0\ 0\text{-}1\ 1)^T = Sgn(2\ \ 2\ \text{-}1\ 1)^T = |13\rangle \tag{17}$$

(3) three missing bits

$$[13]\ (0\ 0\ 0\ \ 1)^T = Sgn(1\ 1\text{-}1\ 0)^T = |12\rangle$$

$$[13]^2(0\ 0\ 0\ 1)^T = Sgn(1\ 1\ \text{-}1\ 3)^T = |13\rangle \tag{18}$$

(4) four missing bits

$$[13]\ (0\ 0\ 0\ 0)^T = Sgn\ (0\ 0\ 0\ 0)^T = (\text{-}1\text{-}1\text{-}1\text{-}1)^T = |0\rangle$$

$$[13]^2(0\ 0\ 0\ 0)^T = Sgn\ (\text{-}1\text{-}1\ 3\ \text{-}1)^T = |2\rangle$$

$$[13]^3(0\ 0\ 0\ 0)^T = Sgn\ (\text{-}3\text{-}3\text{+}3\text{-}3)^T = |2\rangle \tag{19}$$

which converges to a fixed point that is precisely the bipolar complement to $|13\rangle$. In other words, the phase information is lost as an overall minus sign in the last case.

The following sequence of reversing successively from the bipolar bits illustrate tolerance of erroneous bits.

(1) one erroneous bit.

$$[13](-1\ 1\text{-}1\ 1)^T = \mathrm{Sgn}(3\ 1\text{-}1\ 1)^T = |13\rangle \qquad (20)$$

(2) two erroneous bits.

$$[13](-1\ -1\ -1\ 1)^T = \mathrm{Sgn}(1\ 1\ 1\ -1)^T = |14\rangle$$

$$[13]^2(-1\ -1\ -1\ 1)^T = \mathrm{Sgn}(-1\ -1\ -1\ 1)^T = |1\rangle$$

$$[13]^3(-1\ -1\ -1\ 1)^T = \mathrm{Sgn}(1\ 1\ 1\ 1)^T = |15\rangle$$

$$[13]^4(-1\ -1\ -1\ 1)^T = \mathrm{Sgn}(1\ 1\text{-}3\ 1)^T = |13\rangle \qquad (21)$$

(3) three erroneous bits.

$$[13](-1\ -1\ -1\ 1)^T = \mathrm{Sgn}(-1\ -1\ 1\ -3)^T = |2\rangle \qquad (22)$$

which also converges to a fixed point that is also the bipolar complement of $|13\rangle$.

**Generalization within a subspace**

We consider the ability to recognize a new vector that is different from the stored vec In other words, an AM can recognize its related vectors that has not been memorized before. recognition, we mean convergence to a different fixed point. In this sense, we say that the AM c generalize its memory to include other fixed points.

In the case of bipolar vectors, if and only if a new vector **x** is orthogonal to the stor vectors, associative recall "converges in a cycle of two" as defined in the following iterations:

$$\mathrm{Sgn}([T]\,|x\rangle) = -|x\rangle \qquad (23a)$$

$$\mathrm{Sgn}(-[T]\,|x\rangle) = +|x\rangle \qquad (23b)$$

This necessary and sufficient condition allows us to determine efficiently the orthogonali between a new vector and all the stored vectors.

We shall show that when a new vector $|11\rangle$ is presented to the AM [13], due to t orthogonality between $|13\rangle$ and $|11\rangle$ and traceless property of [13],

$$[13]\,|11\rangle = \mathrm{Sgn}(-|11\rangle) = |4\rangle, \text{ and}$$

$$[13]^2|11\rangle = |11\rangle \qquad (24)$$

Once the system has acknowledged the second vector $|11\rangle$, it is incorporated into the matrix storage.

$$4[T_2] = [13,11] = [13] + [11]$$
$$= |13\rangle\langle 13| + |11\rangle\langle 11| - 2\delta \quad (25)$$

If another vector, $|7\rangle$ is presented,

$$[13,11]\,|7\rangle = \mathrm{Sgn}(-2\,|7\rangle) = |8\rangle, \text{ and}$$

$$[13,11]^2\,|7\rangle = \mathrm{Sgn}(4\,|7\rangle) = |7\rangle \quad (26)$$

Thus, we enlarge the memory storage to have three memorized states.

$$4[T_3] = [13,11,7] = [13] + [11] + [7] =$$
$$|13\rangle\langle 13| + |11\rangle\langle 11| + |7\rangle\langle 7| - 3\delta \quad (27)$$

This process is continued until the 4-bit orthogonal subspace (p=3) is filled up.

$$4[T_4] = [13] + [11] + [7] + [14] \quad (28)$$

We have demonstrated the ability to include other orthogonal vectors that have not been stored before. This example also shows the important consequence of traceless storage through its contribution to the "generalization by interpolation within the orthogonal subspace".

Given a table of orthogonal vectors, one may argue that computing inner products will also determine orthogonality. However, inner products must be done pairwise among all vectors and become inefficient as the number of vectors gets large. The above method remains efficient for all sizes.

One may furthermore argue that the difficulty is not how to construct orthogonal set, but to select critical bipolar features from gray-scale, imperfect images.

**Algorithms for Construct A Critical Feature :**

We shall not rely on the auto-AM to select features. One can carry out one's favorite image processing procedure to extract a set of gray-scale feature vectors, $\{|F\rangle\}$. Bipolar feature vectors are preferred in AM because of demonstrated fault-tolerance and the special ability of traceless outer product that allow a quick convergence to a fixed point of cycle two. Given a gray-scale feature vector $|F\rangle$, several procedures for generating a bipolar feature vector are given. The first procedure is "bipolarization", i.e. ,

$$|f\rangle = \mathrm{Sgn}(|F\rangle - \text{threshold}) \quad (29)$$

The second procedure is to use the Walsh transform. We apply two-dimensional W transform (as orthogonal bipolar vector space{ $|w_i>$ }) to all gray-scale features. We select bipolar feature vector from a specific Walsh base vector that is associated with the maxim coefficient in the Walsh transform.

$$|f> = \mathrm{Sgn}(\mathrm{Max}_j\ (\Sigma\ |w_i><w_i|F>) - \text{threshold}) \qquad (30)$$

where the orthonormality condition of Walsh base vectors is inserted to relate to the first meth

$$\Sigma\ |w_i><w_i\ | = [1] \qquad (31)$$

The third and the fourth procedures are to extract from the arbitrary feature vector | the closest vector $|g>$ from either the bipolar orthogonal feature set {$|N>$ } or the { $|F>$ } using following traceless associative memory storage.

$$|g> = \mathrm{Sgn}(\ [\Sigma\Sigma|N><\ F|]\ |\ G> - \text{threshold}) \qquad (32)$$

$$|g> = \mathrm{Sgn}\ (\ \Sigma\ c_F\ [\ |F><F\ ]|\ G> \ - \text{threshold}) \qquad (33)$$

The linear combination coefficients { $c_F$ } may be determined by the statistics of **sin character distortions and variances** (similar to finding the normal modes that diagonalizes covariance matrix and the **Karhunen-Loeve** orthogonal procedure used for outer prod representation of 2-D imagery). Furthermore, the statistics of **character pair distortions, suc two scripted zeros,** could be used to determine the coefficients so as to resolve the proble recognizing ***connected character*** and ***broken character*** after segmentation. We will not go i details in this approach, because of its problem-dependent nature.

The mechanism to select critical features is given as follows.

(1) Human being picks a critical feature (pictures) among the set of distorted, handwrit characters, e. g. the extra stroke among *O, P, Q* .

(2) Walsh transform the selected feature.

(3) Pick the Walsh function that has the largest transform value.

We choose a feature vector that is closest to the Walsh vector associated with the largest Wa transform coefficient, and the rest follows from the procedure described in eq (24-28). We call s a set of features the critical features.

**Lessons to be learned about applying associative memory to pattern recognition:**

AM can only do so much. There is no way to judge the correctness of an associative re except by the convergence to a fixed point. One can only assign meaning to those fixed poi whether it is new or old. The proven capabilities of the AM model are (1) missing and errone

bits recovery, and (2) the creation of new orthogonal vectors, as illustrated above. Therefore, to apply AM to pattern recognition, one must apply human interpretations to those capabilities.

Since learning is by trial and error, it is a continuous process. Suppose that a feature vector with many components representing many features (such as leg-feature and fur-feature, etc, for a tiger, coded fully as $|13\rangle$) has been memorized by the traceless outer product. Furthermore, suppose that only certain features are known in a sequence of imperfect input vectors. (I. e., some feature values are missing. e. g. , the first in the sequence is (0, 0, 1, 1)). Then, the AM can fill in the missing bits. After three iterations, one finds $(-1, -1, 1, -1)^T = |2\rangle$. One can then enlarge the traceless outer product memory to include both vectors, [13, 2]. One examines the second input vectors (0, 0, 1, 1). One can verify that the enlarge memory can indeed recall the vector $|2\rangle$, which correspond to, say, a lady, rather than a tiger. The AM "mental" capacity of recognizing other distinct objects when they show up has been demonstrated. Following this line of thought, the different subspace of different size could be assigned for different classes of objects related by a hetero-associative memory of a rectangular matrix. Such a recognition of different classes requires a complete feature set coded in the AM. It can fill all orthogonal subspaces by the "generalization procedure" illustrated in Eq(24-28).

## 3. ATTENTIVE ASSOCIATIVE MEMORY

Recently, Amari et al has studied the dynamics of such a system, which we will give a simple theorem. We summarize our model equations as follows:

$$\langle n | m \rangle \equiv B \delta_{n,m} \tag{34}$$

$$[T] |n\rangle = \lambda_n |n\rangle \tag{35}$$

The simple model of attentive associative memory $[T]$ is a linear combination of outer products based on the set of orthogonal feature vectors, $\{ |n\rangle , n = 1, \ldots M\}$, and a cue of initial state $|Q\rangle$ that determines the set of attention parameters $\{ a_n \}$ as follows:

$$a_n = \langle n | n \rangle - \langle n | Q \rangle \tag{36}$$

$$B [T]_{ij} = \Sigma^M_{n=1} a_n |n_i\rangle\langle n_j| - A [\delta_{i,j}] \tag{37}$$

that is traceless, $Tr\, \delta_{i,j} = Tr\, |n_i\rangle\langle n_j| = B$, giving

$$A \equiv \Sigma^M_{n=1} a_n \tag{38}$$

and

$$\lambda_n = a_n - (A/B) \tag{39}$$

The attentive memory capacity A and eigenvalue $\lambda_n$ are reduced to Hopfield's memory capacity M and a degenerate eigenvalue $\lambda$, in case of a uniform attention( i.e. $a_n = 1$),

$$\lambda \equiv 1 - r \tag{40}$$

where Amari's pattern ratio $r \equiv (M/B)$ is defined for M bipolar words (states) of B bits (neurons) each.

The dynamics is assumed to be governed by matrix-vector inner product

$$Q(t+1) \equiv Sgn( [T] Q(t) ) \tag{41}$$

where a point nonlinear ity function is defined as $Sgn(x) = +1$ if $x > 0$, and $-1$ if $x < 0$. succesive associative recall gives the iteration, indexed by $t = 0, 1, 2, \ldots$, such that $Q(t) = Q$ when 0.The eigenvalue spectrum, not the distance alone, is a proper macroscopic parameter to explain transient dynamical behaviors of the recalling process. In particular, the direction cosine

$$S_m(t) ) \equiv < m | Q(t) > / < m | m > \tag{42}$$

has been derived and the logarithmic derivative is given by

$$(d/dt) \log ( 1 - S_m(t)) < \log ( \lambda_m / 2 ) < 0 \tag{43}$$

Convergence to a specific m-th state is guaranteed if m-th eigenvalue ( $\lambda_m$ ) is bounded $2 > \lambda_m >$

**Theorem 1** about the lower bound says that paying attention (i.e. non-uniform $a_n$ always increases the memory capacity A ) $\Sigma^M_{n=1} a_n > M$ with a faster convergence proportional to the eigenvalue $\lambda_m > \lambda \equiv 1 - r$

We conjecture that the statistical neurodynamics of associative memory may have similar beha to the deterministic dynamics of attentive associative memory with a non-white eigenva spectrum due to random initial conditions that change with respect to the initial guess vector | >, t =0. The difference vector between | Q(t) > from | m > has an inner product norm defined as

$$2 D_m(t) \equiv < m | m > - < m | Q(t) > \tag{44}$$

If we assume that paying attention to the initial small guess error $2 D_m(0)$ amounts to choos nonuniform and biased storage

$$a_m = 2 D_m(0) \geq 1 \tag{45}$$

and all other coefficients to be identical to 1

$$a_n = 1 \quad , n \neq m. \tag{46}$$

By definition

$$A = M + 2\,D_m(0) - 1. \tag{47}$$

**Theorem 2** about the upper bound of $\lambda_m$ assumes that if a small difference vector between the input $|Q>$ and the specific state $|m>$, is used as the attention parameter $a_m$, Eq(31a), then the critical relationship between the Amari's pattern ratio r and the initial error is analytically found for successful recalls.

$$2\,D_m(0) < 2 + (M + 1)/(B - 1) \tag{48}$$

The maximum permissible Hamming distance $D_H$, from the desired **m**-th state to be reached after iterative recalls, is given by the formula

$$D_H \leq (B/2) - 1 - [(M - 1)/2(B + 1)]((B/2) - 1 - (r/2)) \tag{49}$$

## 4. Conclusion

Associative memory (AM) works like a match filter , but does so efficiently. It should not be applied to image domain directly. Rather, it should be applied to feature domain so that a relatively small AM can do useful tasks at the present technology.

We shall not rely on the auto-AM to select features. Instead, features should be selected using human judgement. However, auto-AM will help us find critical features and hetero-associative memory can perform feature extraction efficiently.

There exists a large body of knowledge pertaining to features selection and extraction and pattern classification for traditional optical character recognition in the literature. This body of knowledge should be tapped and coupled with associative memory. One should not rule out the use of traditional classification techniques (such as syntactical) as extraction of high-level features which then become part of the input feature vector to an AM.

Classical pattern recognition has been demonstrated with a relatively greater success in machine-printed character recognition compared to handprinted character recognition. Difficulty may be rooted in the lack of generalization and abstraction due to machine's limited one-dimensional knowledge representation. In principle, AM should be able to complement traditional OCR with 2-D knowledge representation. Various degrees of abstraction can be achieved through a multi-layer, two-dimensional AM architecture. Note that the present technology has evolved to the point where 2-D memory (chip or optical disk) is not more expensive than 1-D memory storage with logic unit tree content addressable memory processor.

In conclusion, we can combine traditional wisdom in traditional OCR with simple implementable in present technology to form a human-intelligence-endowed neu network.

Character segmentation is an important step in character recognition. Fukushima h developed neural network model (selective attention) for character segmentation in h Neocognitron [Fukushima (1987)]. The attentive associative memory model implement opto-electronically by Athale, Szu & Friedlander (1986) can be augmented by a priori probabili compiled by a character-pair correlation function of connected characters. This is an interesti area for more research.

Inputs to associative memory are linear vectors whereas inputs to OCR are rectangul arrays. Can associative memory replicate the concept of (2-D) neighborhood? The tw dimensional transform that preserves the neighborhood relationship should be used for ima pre-processing before applying AM to the pattern. For example, 2-D Walsh transform can gi a 1-D base Walsh vector (associated with the largest coefficient) as input feature vector to t AM.

Can AM perform syntactical parsing [Ali and Pavlidis (1977)] or rule-based structu analysis [D'Amato (1982)]? Any traditional classification technique can be used to extract hig level features for AM.

How can AM extract position and rotation invariant features? [cf. Szu (1986), Messner a Szu (1987)].

One difficulty in applying backpropagation network has been network size-scali problem. One way to circumvent it has been to extract a small number of features as input. Burr (1987), Gullichsen and Chang (1987)]. Recent advances by Ballard in 1987 permit parti connectivity between two successive layers which avoids combinatorial explosions oft encountered when the input layer is directly connected to image pixels. Thus, spatial patte relationship can be efficiently preserved in such a network while coarse-graining betwe successive layers can desensitize pattern variation in input images.

An AI extension of the simple AM model is attentive associative memory, (AAM), th allows us to apply AI to pay a non-uniform attention to each term of outer product storage, i a linear combination of outer products in which the set of combination coefficients determined by AI rule-based system, e.g. the frequency distribution of distorted characte according to user group profiles, e.g. left hand writing versus righthand writing. The efficien of the closed system of rule-based knowledge representation of AI using the tuple storage combined with the flexibility of the non-rule based open system using the matrix knowled representation of NI ( coined for either neural, or network, or natural intelligence). Thus, t ability of generalization and abstraction becomes possible for AI, and is demonstrated in combined intelligent system of AI & NI. We can endow a simple neural network architectu based on a small set of neurons with a human-like intelligence by combining the tradition rule-based AI wisdom with non-rule-based learning. This is achievable because OCR requir

better feature vectors obtained from other discipline in the sense of fault tolerance that neural networks built at the present technology can already provide with.

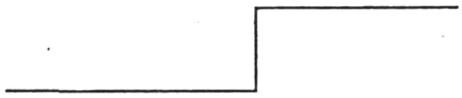
## Appendix: Generic Definition of Neural Networks

Associative memory is a special model of neural networks. Examples of associative recalls from partial images and the success of nonlinear signal processing are recorded in the literature [cf. Kohonen (1984)]. An axiomatic definition is outlined as follows.

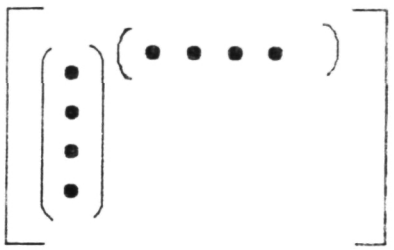
We shall define **three kinds of neurons**: fine-grained, medium-grained and large-grained processor elements (PEs). A fine-grained PE, represented by the lower case word ***neuron*** , has no internal memory analogous to neurons in the hippocampus part of the brain that is responsible for fault-tolerant associative recall. A medium-grained PE, ***Neuron,*** has a built-in memory analogous to Neurons in biological sensory and motor control which are responsible for reactions to approaching danger. A large-grained PE, ***NEURON*** , has built-in memory, control logic, and communication capabilities equivalent to a computer. NEURONs occur in nature in the form of grandmother cells or pacer/conductor cells.

These three types of neurons and their associated circuits have **four kinds of interactions**: ***(1) exciting, (2) inhibiting, (3) bursting, (4) grading and delaying transmission.*** In general they follow the law of the middle response or the sigmoid function (hyperbolic tangent or logistic functions) to amplify weak signals with a nonlinear quick rising function and suppress strong signals with a nonlinear tapering off saturation function. The generic definition of a Neural Network is a system which is:

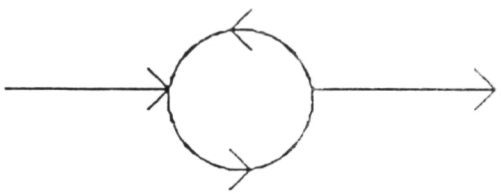
1. ***Non-linear*** ≈ sigmoid function ≈ point non-linearity (hard limiting) shown as follows:



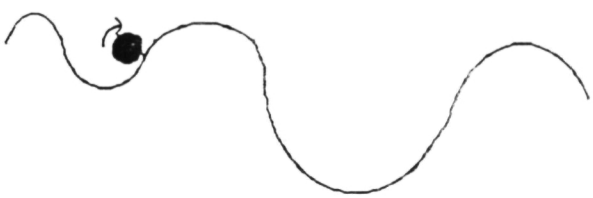
2. ***Non-local*** ≈ weighted outer product ≈ outer product (white spectrum) shown as follows:



3. *Non-stationary* ≈ piecewise time stationary ≈ iterative algorithm shown as follows:



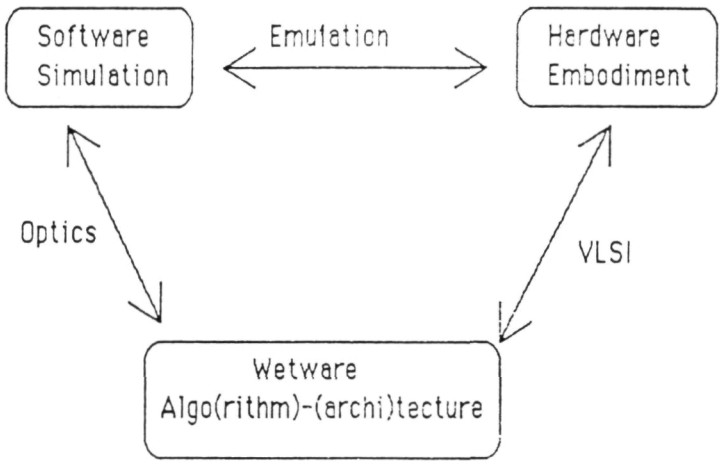
4. *Non-convex* ≈ constrained global optimization ≈ simulated annealing schematica shown as follows:



5. *Other attributes yet to be discovered* .

These successive approximations of the four *non*--principles, indicated by wiggly equal signs in (1-4), makes possible the unveiling of the complex and nonlinear neural (bra behavior. This is possible with the use of powerful computers and more accurate models intelligent functions. The theory is amenable to numerical simulations due to ***piecew linear, regionally local, temporarily stationary, and locally convex approximations.***

Three decades ago, Rosenblatt and co-workers built the **perceptron** solely based upon t first attribute (nonlinearity) with stochastic implementations. Thus, with hindsight, it was r surprising that Minsky and Papert could show a limited utility and propose useful alternati **artificial intelligence (AI) rule- based systems.** ***AI works in closed systems where rules gove while neural intelligence (NI) works in open systems where rules have yet to be discover*** Various exploitation of these efforts in neural networks are:




The term wet-ware, coined by Carver Mead, is neither software nor hardware, but more like a Hecht-Nielsen's net-ware based on non-programmable but trainable networks. A special version of layered neural networks has been demonstrated with the ability of phonetic interpolation in the Rumelhart, Sejnowski connectionist's networks, such as Net-Talk, Boltzmann and Cauchy Machines, and error back propagation networks.

**Acknowledgement**

The work has been supported by ONR under IST/SDIO program. Discussions with John Tan and Frank Polkinghorn are gratefully acknowledged.

## 6. REFERENCES

Ahmed, P. and Suen, C., "*Computer recognition of totally unconstrained handwritten ZIP Codes*," International Journal of Pattern Recognition and Artificial Intelligence,Vol. 1 (1987), pp.1-15.

Ali, F. and Pavlidis, T., "*Syntactic recognition of handwritten numerals*," IEEE Transactions on SMC, Vol. 7 (1977), pp.537-541.

Athale, R.A., Szu, H.H., & Friedlander, C.B., "*Optical implementation of associative memory with controlled nonlinearity in the correlation domain*," Optics Letters, Vol. 11 (1986), pp. 482-484

Burr, D., "*Designing a handwriting reader*," IEEE Trans. on PAMI, Vol.5 (1983)

Burr, D., "*Experiments with a connectionist text reader*," Proc. of IEEE International Conference on Neural Networks, 1987, Vol. IV, pp. 717-724.

D'Amato, D., et al, "*High speed pattern recognition system for alphanumeric handprinted characters*," Proc. of Pattern Recognition and Image Processing, 1982, pp.165-171.

Duda, R. and Hart, P., *Pattern Classification and Scene Analysis* Wiley-Interscience, 1973.

Duerr, B., Haettich, W., Tropf, H. and Winkler, G., "*A combination of statistical and syntacti pattern recognitionapplied to classification of unconstrained handwritten numerals,*" Patte Recognition, Vol. 12 (1980), pp.189-199.

Fukushima, K. and Miyake, S., "*Neocognitron: a new algorithm for pattern recognition tolerant deformations and shifts in position,*" Pattern Recognition, VOL. 15, 1982, pp. 455-469.

Fukushima, K., "*A neural network model for selective attention,*" in Proc. of IEEE Internatio Conference on Neural Networks,1987, Vol. II, pp. 11-18.

Gullichsen, e. and Chang, E., "*Pattern classification by neural network: an experiment system icon recognition,*" Proc. of IEEE International Conference on Neural Networks, 1987, Vol. I pp.725-732.

Hull, J., et al., *Optical Character Recognition Techniques in Mail Sorting: A Review of Algorithn* Technical Report 214, State University of New York at Buffalo, Department of Computer Scien 1984.

Kohonen, T.,"*Self-Organization and Associative Memory,*"Springer-Verlag, 1984.

Stringa, L., "*LCD: a formal language for constraint)free hand-printed character recognition,*" P of International Conference on Pattern Recognition, 1978, pp. 354-358.

Schurmann, J., "*Multifont word recognition system with application to postal address readin* Proc. of Int. Conf. on Pattern Recognition, 1976, pp. 658-662.

Schurmann, J., "*Reading machines*", Proc. of Int. Conf. on Pattern Recognition, 1982, pp.1031-104

Shridhar, M. and Badreldin, A., "*Recognition of isolated and simply connected handwrit numerals,*" Pattern Recognition, Vol.19 (1986), pp.1-12.

Suen, C. Y., Berthod, M. and Mori, S., "*Automatic recognition of handprinted characters - the st of the art*", Proceedings of the IEEE, vol. 68 (1980), pp. 469-487.

Szu, H. H., Caulfield, H.J., "*Optical Expert Systems,*" Applied Optics, Vol. 26, pp. 1943-1947, 1987

Szu, Harold H., "*Three layers of vector outer product neuural networks for optical patte recognition.*", In H.Szu (Ed.) *Optical and Hybrid computing* (1986) (pp. 312-330), Bellingham, W Society of Photo-Optical Instrumentation Engineers.

Szu, H. H. & Messner, R. A. ,"*Adaptive Invariant Novelty Filters,*" Proceedings of IEEE, Vol. (1986), pp. 518-519

Szu, Harold H., "*Globally connected network models for computing using fine-grained processing elements*," In C. P. Wang (ed.) Proceedings of International Conference on LASERS '85, pp. 92-97, Society for Optical & Quantum Electronics, P.O. Box 245, McLean VA 22101

Szu, Harold H., and Tan, John,"Can associative memory recognize charactors ?", Third Advanced Technology Conference, U.S. Postal Service, Washington D.C. May 3-5, 1988

Tou, J., and Gonzalez, R., *Pattern Recognition Principles*, Addison-Wesley, 1974.

Winston, P. H. , "*Artificial Intelligence*," Reading , Mass. Addison-Wesley 1984, 2-ed.

# NOTES