Accepted for publication in the
IEEE Transactions on Neural Networks
(Date of acceptance January 1997)

CHARACTERISTICS OF MULTIDIMENSIONAL HOLOGRAPHIC ASSOCIATIVE MEMORY IN RETRIEVAL WITH DYNAMICALLY LOCALIZABLE ATTENTION

Javed I. Khan & D. Y. Yun

Contact:
Laboratories of Intelligent and Parallel Systems
Department of Electrical Engineering
493 Holmes Hall, 2540 Dole Street
University of Hawaii at Manoa
HI-96822, USA

Phone: (808)-956-3868
Fax: (808)-941-1399
javed@hawaii.edu

SUMMARY

This paper presents the performance analysis (capacity and retrieval accuracy) of Multidimensional Holographic Associative Memory (MHAC). MHAC has the unique ability to retrieve pattern-associations with changeable attention. *In attention actuated retrieval* the user can dynamically select any subset of the elements in the example query pattern and expect the memory to confine its associative match only within the specified field of attention. Existing *artificial associative memories* lack this ability. Also most of these models need at least 50% of bits in the input pattern to be correct for successful retrieval. MHAC, with the unique ability of localizable attention, can retrieve information correctly even with cues as small as 10% of the query frame. This paper investigates the performance of MHAC in attention actuated retrieval both analytically and experimentally. Besides confirmation, the experiments also identify an operational range space (ORS) for this memory within which various attention based applications can be built with a performance guarantee.

CHARACTERISTICS OF MULTIDIMENSIONAL HOLOGRAPHIC ASSOCIATIVE MEMORY IN RETRIEVAL WITH DYNAMICALLY LOCALIZABLE ATTENTION

Javed I. Khan & D. Y. Yun

Laboratories of Intelligent and Parallel Systems
Electrical Engineering Department
University of Hawaii at Manoa
javed@hawaii.edu

ABSTRACT

This paper presents the performance analysis (capacity and retrieval accuracy) of Multidimensional Holographic Associative Memory (MHAC). MHAC has the unique ability to retrieve pattern-associations with changeable attention. In attention actuated retrieval the user can dynamically select any subset of the elements in the example query pattern and expect the memory to confine its associative match only within the specified field of attention. Existing artificial associative memories lack this ability. Also most of these models need at least 50% of bits in the input pattern to be correct for successful retrieval. MHAC, with the unique ability of localizable attention, can retrieve information correctly even with cues as small as 10% of the query frame. This paper investigates the performance of MHAC in attention actuated retrieval both analytically and experimentally. Besides confirmation, the experiments also identify an operational range space (ORS) for this memory within which various attention based applications can be built with a performance guarantee.

1. INTRODUCTION

The modern research in distributed and parallel models of Artificial Associative Memory (AAM) started with the McCulloch and Pitts' invention of *formal neuron* in 1943. This invention for the first time provided a formal architecture for a brain like distributed processing of information. It was extraordinary. Because, reinforced by the theory of *symbolic logic* (Russell & Whitehead, 1910, 1912, 1913), it promised universal computability and artificial realizability of almost unlimited complex systems [21,24]. The optimism it sparked was followed by a vigorous and immensely productive era of research in artificial neuro-computing.

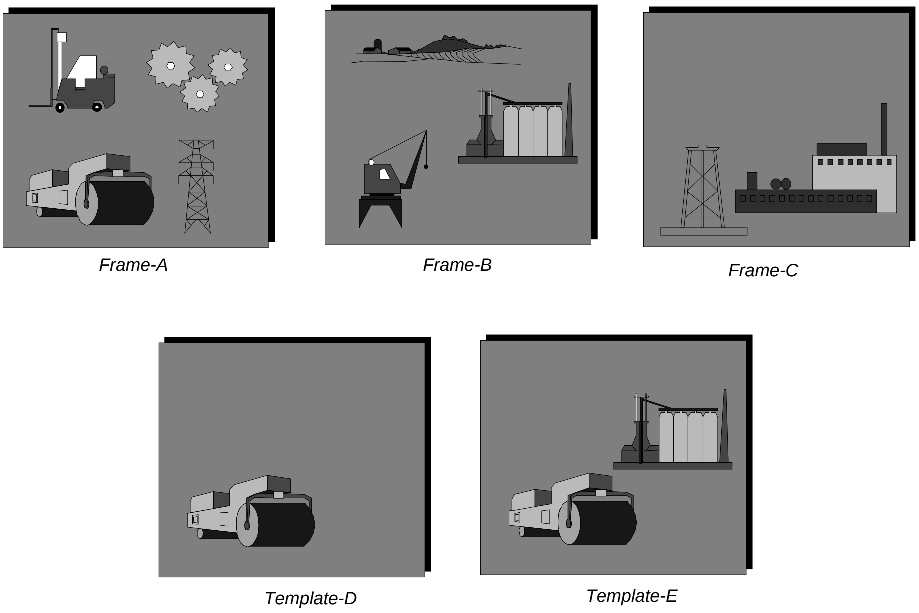
However, beginning with Rosenblatt, till today researchers have focused, and in many ways confined themselves to the perfection of the learning behavior of these artificial systems. During these years, increasingly more intricate and complex properties of learning phenomena have been pursued in great depth. Versatility (how arbitrary complex associations can be learned), efficiency (how more patterns can be learned), learnability of causality (Klopf 1987), learnability of temporal relations, learning in continuum of time (Grossberg 1967), self-organization (Kohonen 1987, Oja 1982), autonomous unsupervised adaptation (Grossberg 1976, Carpenter & Grossberg, 1987) are just a few examples of the successes and intricacies through which research in artificial learning matured [6,15,16,22,7,2]. Surprisingly, during these enormously productive years, few attempts had been made to examine the recollection aspects of AAMs other than assuming a very simple model of retrieval for all these forms of learning.

Almost all of the proposed learning models since McCulloch and Pitts have been constructed on the assumption of a simple and restricted retrieval scenario[1]. In this scenario the sample from the content which is used during query is a close replica of the target. However, more complex and versatile retrieval formalism is not only conceivable but also seems to be an integral part of natural associative memories. The ability to almost effortlessly infuse attention during retrieval is one such aspect of natural recollection.

The phenomenon is explained through an example of image perception. Let an associative memory be allowed to learn the image frames A, B and C of Fig-1. If during the retrieval, template-D is used as a sensory input, then it is natural to expect that the memory should retrieve frame-A based on the roller. It appears to be the most cognitively significant index in the template. However, it can be demonstrated that most of the conventional AAMs

[1] Consequentially, most of these learning methods break down when the test of learning is based on the generalized retrieval scenario.

instead, will retrieve frame-C as the closest match (indeed B and C are closer to D than A; both in *least mean square* (LMS), and *maximum dot-product* sense). The reason for such an unexpected result is the statistical weakness of the cognitively important roller pixels compared to the statistical strength of cognitively less important background pixels. In contrast, a natural memory seems to be immune to such statistical weakness and can retrieve information by localizing attention on cognitively important zones.


Frame-A

Frame-B

Frame-C

Template-D

Template-E

Fig-1 Attention Modulated Retrieval

The most intriguing aspect of natural associative memory is that it can change the distribution of attention over element space dynamically during query. Consider template-E. There are two objects of focus and two possible answers. If desired, a natural memory can shift its attention to any other object (for example, on the Plant) in the template and retrieve entirely different match (frame-B) apparently without any significant internal reorganization. In contrast, a conventional AAM lacks such flexibility. For a given state of learning, it acts as a deterministic machine where each initial state flows into a pre-determined single attractor. Conventional AAMs have no mechanism to accommodate dynamic (post-learning) change in the distribution of attention over its element space.

A serious consequence of such attention deficiency of conventional AAMs is their inability to work with a small cue. A conventional AAM requires the effective cue to be statistically significant compared to the overall pattern size. For correct retrieval, the effective cue should be at least 50% of the pattern size for any AAM [10]. This is quite unrealistic for many applications. Interestingly, experiments performed by previous researchers contain empirical evidence of such severe retrieval inadequacy of the existing AAM models [18, 27,8].

Khan [10] has recently demonstrated that an associative computation model called *Multidimensional Holographic Associative Computing* (MHAC), based on hyperspherical representation can overcome these limitations. It has been demonstrated that MHAC is capable of retrieving associatively learnt information with dynamically changeable attention over the element set of query pattern.

The representation, learning and retrieval model of this memory has been derived from the principles of Holography[2]. The detail of the derivation of MHAC from holographic representation and its analyses can be found in [10]. The paper presents the performance analysis of this model. This paper formally investigates the relationship between degree of focus, retrieval accuracy, capacity, and scalability of this attentive memory both analytically and experimentally.

The following section first describes the concept of this attentive memory. Section 3 briefly presents the computational model. Section 4 then presents the detailed analysis of performance of this model. Finally, section 5 presents the result of extensive computer simulation that empirically confirms the analytical derivations. In addi-

2 An excellent background description of holography itself has been recently published in [23].

tion to the empirical observation of the critical characteristics of this memory, this section also presents the *operational range space* (ORS). ORS can assist application designers in developing efficient applications on this model by providing precise value ranges of the critical design parameters.

2. ATTENTIVE MEMORY

2.1 Concept: Bimodal Associative Memory

A pattern is a collection of elements. Let a stimulus and corresponding response patterns be denoted by the symbolic vectors $S^{\mu} = [s_1^{\mu}, s_2^{\mu}, \ \ldots \ , s_n^{\mu}]$and $R^{\mu} = [r_1^{\mu}, \ r_2^{\mu}, \ \ldots \ , r_m^{\mu}]$. Each of the individual elements in these vectors represents a piece of *information*. The superscript refers to the index of the pattern and the subscript refers to the index of the element in it. The values of these elements correspond to a measurement obtained by some physical sensor.

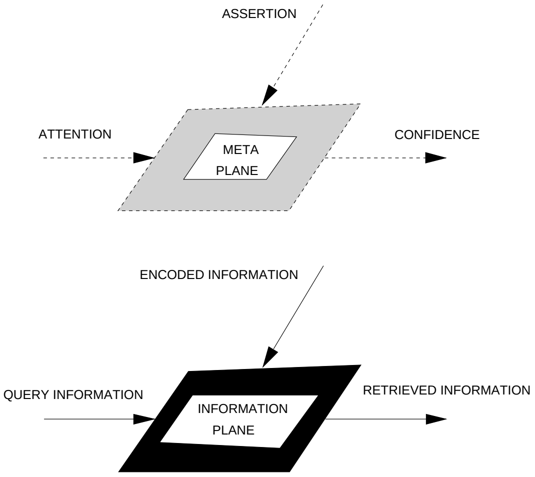


Fig-2 Information Flow Model of Bimodal Memory

A memory has three information channels (as shown in the bottom plane of Fig-2). The first is the encoder input, where stimulus and response pattern pairs are received during learning. The second is the decoder input, where query stimulus pattern is received from inquirer. The third is the decoder output, where the response pattern is generated by the memory as a reaction to the query. A conventional associative memory processes only the above measurement components of information in such a way that:

Definition 1: An ***Associative Memory****, given a set of p stimulus pattern vectors* $\mathbf{S} = \{S^{\mu} \mid 1 \le \mu \le p\}$ *and a set of equal number of response pattern vectors* $\mathbf{R} = \{R^{\mu} \mid 1 \le \mu \le p\}$*, learns the relationship between a stimulus member* $S^{\mu} \in \mathbf{S}$ *and the corresponding response member* $R^{\mu} \in \mathbf{R}$ *in such a way that, given a query pattern* S^Q*, it can retrieve a pattern* $R^R \approx R^T$ *such that* $R^T \in \mathbf{R}$*, and* S^Q *is closest to* $S^T \in \mathbf{S}$ *according to some matching criterion* $D()$*.*

An associative memory system is comprised of (i) a learning algorithm $\mathbf{A}_{\text{learn}}$ which converts all the $\{S^{\mu}, R^{\mu}\}$ associations into some internal representation, (ii) a physical storage medium and representation formalism **AM** to store the associations, (iii) a decoding algorithm $\mathbf{A}_{\text{retrieve}}$ to recollect stored information R^R from a given query stimulus S^Q, and (iv) a matching criteria $D()$ to measure the closeness of stimulus patterns to the query pattern. The actual form of $D()$ may vary between the AM models. In section 2.3 some pertinent forms of this function has been further illustrated.

A conventional memory formalism processes only the measurements associated with the information elements in the above model. In contrast, the conceptual memory model of MHAC is based on a formalism which assumes that the *trust* in each piece of transacted information is inherently nonconforming and measurements associated with the information elements are individually susceptible to distortion, loss, or even purposeful disregard. The formalism includes the *meta-knowledge* about the state of each given piece of information (measurement) as an integral part of its basic notion of information.

The proposed formalism adopts an additional meta-knowledge plane (as shown in the upper plane in Fig-2). The linguistic interpretation of the quantities of this meta plane varies depending on the channel. For the encoded information, this meta-knowledge corresponds to a form of *assertion* from the encoder. For the query pattern, it corresponds to a form of *attention* on the part of inquirer. For the memory response, it corresponds to the *confidence* on the retrieved information as assessed by the memory itself.

Formally each of the elements of information is modeled as a bi-modal pair $s_k^{\mu} = \{\alpha_k^{\mu}, \beta_k^{\mu}\}$. Where α represents the *measurement* of the information elements and β represents the *meta-knowledge* associated with this measurement.

The above formalism, in the context of a general memory (irrespective of its implementation mechanism) which computes on imperfect knowledge, generates some specific expectations about the operational behavior of these meta quantities. These are stated below.

Expectation on the Inflow of Meta Knowledge: The memory matching criterion should put more importance to a piece of information that is attributed with high *degree of inflow β than to a piece attributed with low β in the query. The expectation can be stated as a matching criterion:*

$$D\left(S^Q, S^{T_a}, \mathrm{B}\right) = \sum_{i}^{n} \beta_i dist\left(\alpha_i^Q, \alpha_i^{T_a}\right) \qquad \ldots(1)$$

Here N is the set of all elements in the pattern vector and the set has cardinality n. The index variable i varies from 1 to n and thus the summation includes all the elements in the set N. The function *dist()* denotes a measure of distance between individual pattern elements. The additional input B denotes the meta-vector. From the context of encoding, when B^{μ} is specified dynamically during encoding, this expectation corresponds to a learning criterion that can realizes *learning with changeable assertion* (LCA). From the context of query, when B^{Q} is specified dynamically during query, this expectation corresponds to a matching criterion that can realizes *retrieval with changeable attention* (RCA). It is the later meta-knowledge on which rest of this paper will focus.

The incorporation of meta-knowledge into the basic notion of information goes beyond the important concept of attention. A second symmetric expectation related to the outflow of meta-knowledge provides completeness to this attempt of delineating the behavior of a new memory.

Expectation on the Outflow of Meta Knowledge: If α values of query demonstrate high degree of resemblance to the α values of a priory encoded stimulus pattern, then memory should retrieve the associated response α with higher degree of accuracy and high degree of β. On the other hand, if it does not then it should generate a response with low degree of β, as detailed in Table-1.

Query		Response	
β_{query}	α_{query}	$\beta_{response}$	$\alpha_{response}$
HIGH	*CLOSE*	*HIGH*	*CLOSER*
LOW	*CLOSE*	*HIGHER| LOW*	*CLOSER*
HIGH	*NOT CLOSE*	*LOW*	*CLOSEST| DON'T CARE*
LOW	*NOT CLOSE*	*LOWER*	*CLOSEST| DON'T CARE*

Table-1 Expectations

Inflow Expectation relates to the inward communication of the meta-knowledge into the memory system. An external querying system (it can be a human user or another computer system) supplies the stimulus elements and the additional significance level of each stimulus element. Outflow Expectation relates to the outward communication from the memory to the external querying system. In the reply, the querying system is given back not only the retrieved measurements but also the meta-knowledge *confidence* about the status of the retrieved content. Both of the transfers are essential in the context of imperfect knowledge transaction.

The above expectations essentially constitute the behavioral definition of a memory system which incorporates possibility of imperfection in the given measurements. In the rest of this paper such a memory will be referred to as *Attentive Memory*. The next section presents the actual computational model which can realize at least one instance of the attentive memory by satisfying these expectations.

2.2 Dynamically Changeable Attention

In the context of the above definition of memory the concept of dynamic attention will now be clarified.

Definition 2: ***Attention*** *refers to the fact that any subset[3] $F^Q \in N$ of the elements in the example query pattern S^Q can be specified at post-learning stage as a* ***field of attention*** *and the memory can confine its associative match (by a suitable matching criteria D) only within F^Q.*

One of the most important aspects of attention based retrieval is the dynamic specifiablity of the field of attention. Here dynamism refers to the post-learning changeability of the distribution of attention during query.

If a specific distribution of attention is given during encoding at pre-learning stage, a conventional AAM can hard-encode it in the learned synaptic weights. However, once the learning is over, it does not allow the distribution of attention to be recast during query. For a given learning, it acts as a deterministic machine where each initial state flows into a pre-determined single attractor. Conventional AAMs have no mechanism to accommodate post-learning change in the distribution of attention. Dynamic attention is equivalent to the capability of accommodating varied perspectives on the query pattern.

2.3 Definitions of RCA Queries

The ability of retrieval with changeable attention is reflected by the type of distance evaluation criteria used

[3] The membership in the attention subset can be bivalued or analog. In analog membership case a particular element can be partial members in more than one subsets, provided that all memberships adds up to one.

by any memory. In general it is possible to define the following three matching criteria and corresponding RCA query types for the memory system model defined here.

Definition 3: An ***Unary Attention AM*** *(RCA type-U) is one which retrieves a pattern $R^R \cong R^{T_u}$, such that the distance between its associated stimulus and the query pattern is evaluated by a matching criterion of the form:*

$$D\left(S^Q, S^{T_u}\right) = \min_{1 \le \mu \le p}\left[\sum_{i}^{n} dist(s_i^Q, s_i^\mu)\right] \qquad \ldots(2)$$

This definition corresponds to a matching criterion which considers all elements in the query pattern to be equally important from the searchers frame of reference. It converges on one of the *p* previously learned patterns.

If $F^Q \subseteq N$ represents a subspace of the total element space *N*, then problem that an associative memory can retrieval with <u>changeable attention</u> can be stated in the following form:

Definition 4: A ***Binary Attention AM*** *(RCA type-B) is one which retrieves a pattern $R^R \cong R^{T_b}$, where the set of elements in an attention vector F^Q is dynamically specifiable during query, and the distance between it's associated stimulus and the query pattern is evaluated by a matching criterion of the form:*

$$D\left(S^Q, S^{T_b}, F^Q\right) = \min_{1 \le \mu \le p}\left[\sum_{i}^{F^Q} dist(s_i^Q, s_i^\mu)\right] \qquad \ldots(3)$$

The above retrieval can be further generalized when the attention on a specific element is allowed to be partial. This generalized form of retrieval characterized with <u>changeable analog attention</u> can be stated as follows:

Definition 5: An ***Analog Attention AM*** *(RCA type-A) is one which retrieves a pattern $R^R \cong R^{T_a}$, where the analog attention on the stimulus elements is represented by the dynamically specifiable query vector, $\Lambda^Q = [\lambda_1^Q, \lambda_2^Q, \ \ldots \ , \lambda_n^Q]$, and $0 \le \lambda_i^Q \le 1$. The distance between it's associated stimulus and the query pattern is evaluated by a matching criterion of the form:*

$$D\left(S^Q, S^{T_a}, \Lambda^Q\right) = \min_{1 \le \mu \le p}\left[\sum_{i}^{n} \lambda_i^Q dist(s_i^Q, s_i^\mu)\right] \qquad \ldots(4)$$

Here the analog attention on the stimulus elements is represented by an additional query vector of length n, $\Lambda^Q = [\lambda_1^Q, \lambda_2^Q, \ \ldots \ , \lambda_n^Q]$ where $0 \le \lambda_i^Q \le 1$.

2.4 Retrieval in Current AAM Models

The optimization criteria of the existing neural models directly belong to type-U category. Models which use Hebbian class of learning maximize global dot-product of the patterns [17]. On the other hand, the models which use LMS class of learning maximize global mean square error [29]. There are few other distance measures also (such as *entropy*, *maximum likelihood ratio*, etc.) those have been used as the matching criterion in conventional neuro-computing. However, Hopfield has provided a generalized perspective to analyze the collective behavior of a collection of interconnected neurons irrespective of the specific function they minimize or maximize. He has demonstrated that the convergence (or recollection) behavior of a collection of interconnected neurons can be interpreted as a minimization of some form of energy function [8]. The key features to note in the energy functions of current neural network models are (i) the set operator is a summation process $\mathbf{M} \equiv \Sigma$, and (ii) the scope *N* of the set operator is all inclusive and is based on entire element space, and (iii) the element distance function is only bivariate. These properties of existing neural networks together make them a type-U memory.

Intuitively, the reason that conventional AAMs cannot support dynamic attention is twofold. (a) First, the discrete summation step, which is the foundation stone of *synoptic efficacy rule* (like any other finite summation process), requires almost all its input elements to be present. Although a summing output can tolerate some random statistical distortion of the input values, it can not tolerate selective and deliberate (full of partial) withdrawal of inputs[4]. (b) Secondly, in a scalar (one dimensional) space, it is not possible to create dynamic representation for the notion of 'dont-care'. The meaning of 'dont-care' is equivalent to specifying the state of an element which is not in the attention set F^Q. Any AAM constructed from interconnected cells of such finite discrete integrators (which includes almost all of existing models) suffers from this fundamental limitation. In [11] it has been formally shown that:

Theorem 1: An associative memory constructed by interconnecting cells with the scalar product rule of synaptic transmission specified by the equation below can not realize the retrieval of type-B, or type-A. Where, f() is any single variate function, and s_j is a real valued number in the range I=[0,1], and the weights w_{ij} contains the learned pattern.

$$r = f\left(\sum_{i}^{n} w_{ij}.s_j + b_i\right)$$

A memory based on multidimensional complex representation can overcome the above limitations of conventional AAMs and can support the generalized type-A as well as type-B retrievals.

3. COMPUTATIONAL MODEL

The computational model of the MHAC is conceptually based on optical holography [4,28,23]. The details of this derivation can be found in [10]. This section now briefly describes the model.

3.1 Representation

In this approach, each piece of information is mapped onto a multidimensional complex number (MCN). Each α_k is mapped onto a set of phase elements $\theta_{j,k}$ in the range of $\pi \geq \theta_{j,k} \geq \pi$ through a mapping transformation $m^{+\alpha}(x)$. Corresponding meta information β_k is mapped as its magnitude λ_k through another transform $m^{+\beta}(x)$[5].

$$s_k = \{\alpha_k, \beta_k\} \Rightarrow \lambda_k e^{\left(\sum_{j}^{d-1} \hat{i}_j \theta_{j,k}\right)} \qquad \ldots.(5)$$

Where, each element $s_k(\lambda_k, \theta_{1,k}, \theta_{2,k}, \ldots, \theta_{d-1,k})$ is a vector inside a unit sphere in a d-dimensional spherical space. Each $\theta_{j,k}$ is the spherical projection (or phase component) of the vector along the dimension $\hat{i}_j$. This computational representation will be called *multidimensional complex numeric* (MCN) representation of information.

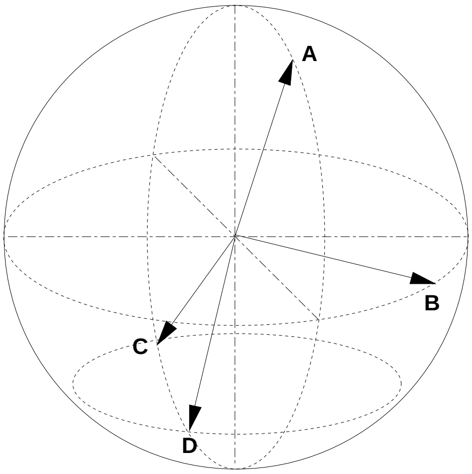


Fig-3 Points on Hypersperical Surface

<u>Mapping of measurements:</u> A class of functions can be used as the mapping transform $m^{+\alpha}()$. The function should be single valued and continuous. For discrete inputs, continuity is required at the defined points. A desirable characteristic of the mapping transform is that it should maximize the symmetry at the phase domain.

<u>Mapping of significance:</u> Any positive valued rule of mapping with the following two constraints can be used as $m^{+\beta}()$. Elements with same magnitude (equi-significant) are required to contribute equally to the subsequent decision stages. An element with magnitude zero should have no effect on the outcome of the computing. In addition, clipping the upper bound of the magnitude to 1.0 establishes a probabilistic interpretation of certain aspects of this representation. If all the elements of a pattern are made equi-significant, this representation becomes functionally equivalent to that of conventional AAMs. However the opportunity to modify these magnitude values dynamically during query provides a new capability of selective attention.

[4] Sherrington's [25] observation on the existence of some form of integration process in the nervous sites is generally used to rationalize the use of linear weighted sum. However still now the theory itself has not been decidedly validated or refuted. More importantly, the weighted average suggested by us does not imply the absence of integration. Sherrington's theory also suggests the existence of temporal summation [9]. Recent evidence suggests that in some cases two neurotransmitters can co-exist in axons. It is also plausible that the pre-synaptic dendrites will also have individual saturations like any other physical channels. All of which can potentially make the summation non-linear even at the channels.

[5] The inverse transformations to revert back from MCN representation are respectively denoted by $m^{-\alpha}()$ and $m^{-\beta}()$.

Combined representation: Thus, each of the information elements is represented as a vector bounded in the unit multidimensional spherical space. A stimulus pattern is computationally represented as:

$$[S^{\mu}] = \left[\lambda_1^{\mu} e^{\left(\sum_j^{d-1} \hat{i}_j \theta^{\mu}_{j,1}\right)}, \lambda_2^{\mu} e^{\left(\sum_j^{d-1} \hat{i}_j \theta^{\mu}_{j,2}\right)}, \ldots . \lambda_n^{\mu} e^{\left(\sum_j^{d-1} \hat{i}_j \theta^{\mu}_{j,n}\right)} \right]$$

A similar mapping on the external scalar response field intensities provides the response representation:

$$[R^{\mu}] = \left[\gamma_1^{\mu} e^{\left(\sum_j^{d-1} \hat{i}_j \phi^{\mu}_{j,1}\right)}, \gamma_2^{\mu} e^{\left(\sum_j^{d-1} \hat{i}_j \phi^{\mu}_{j,2}\right)}, \ldots . \gamma_m^{\mu} e^{\left(\sum_j^{d-1} \hat{i}_j \phi^{\mu}_{j,m}\right)} \right]$$

Here, the phasor ϕ represents the measurement of the retrieved response and γ represents the expected *confidence* (system assigned significance) on ϕ.

History of MCN representation: Use of complex number is not a completely new concept in artificial associative computing, at least in 2-dimension. In 1990 Sutherland [26] in his pioneering work, presented the first truly holographic associative memory with holographic representation and learning algorithm analogous to correlation learning as used here. It is a 2-dimensional special case of the generalized multidimensional phasor representation introduced here. Much of the conventional retrieval based (RCA type-U) characteristics with the 2-dimensional representation of this model have been investigated in depth in this pioneering work. More recently Timothy Masters [20] also reported another 2D-complex valued network with a learning algorithm analogous to Backpropagation. However both of these attempts remained focused on their network's efficiency issues as a conventional adaptive filter (with type-U retrieval). The fundamentally different phenomena of attention (type-A/B retrieval) associated with such representation [11] remained unexplored.

The first artificial system to demonstrate associative phenomena ever, optical holography itself [4], can be considered a complex valued computation mechanism. When pioneering researchers[6], ventured to recreate such fascinating optical transforms artificially on digital computers, they adopted some simplifications to gain efficiency on digital computers. One of those early simplifications was the use of scalar numbers instead of 2D optical wave. All subsequent research adopted this simplified representation. Its implication was hardly ever reinvestigated. In that sense this work is a visit back to the lost dimensionality of representation; and a step beyond. It explores further into a computational model based on multidimensional phasor (instead of only 2-D phasor) representation.

3.2 Encoding

In associative memory, information is stored in the form of associations. In the encoding process, the association between each individual stimulus and its corresponding response is defined in the form of a correlation matrix. This matrix is computed by the inner product of the conjugate transpose of the stimulus and the response vectors, and is stated in equation (6).

$$[X^{\mu}] = [\bar{S}^{\mu}]^T \cdot [R^{\mu}] \qquad \ldots .(6)$$

If the stimulus is a pattern with n elements and the response is a pattern with m elements, then $[X]$ is a $n \times m$ matrix with d-dimensional complex elements.

A suit of associations derived from a set of stimulus and corresponding response is stored in the following correlation matrix X. The resulting memory substrate containing the correlation matrix is referred to as Holograph.

$$[X] = \sum_{\mu}^{p} [X^{\mu}] = \sum_{\mu}^{p} [\bar{S}^{\mu}]^T [R^{\mu}] \qquad \ldots .(7)$$

3.3 Retrieval

During recall, the query stimulus pattern $[S^e]$ is represented by:

$$[S^e] = \left[\lambda_1 e^{\left(\sum_j^{d-1} i_j \theta^{e}_{j,1}\right)}, \lambda_2 e^{\left(\sum_j^{d-1} i_j \theta^{e}_{j,2}\right)}, \ldots, \lambda_n e^{\left(\sum_j^{d-1} i_j \theta^{e}_{j,n}\right)} \right]$$

The decoding operation is performed by computing the inner product between the excitatory stimulus and the correlation matrix X:

$$[R^e] = \frac{1}{c}[S^e] \cdot [X] \qquad \ldots .(8)$$

$$where, \quad c = \sum_{k}^{n} \lambda_k$$

Although the above computation appears analogous to conventional associative computing paradigm, but it displays fundamentally different characteristics than conventional associative computing. They process the measurement component of information quite differently. Next section explains the fundamental distinctions that make this new parallel and distributed computing paradigm capable of supporting type-A & B RCA search.

[6] Following the work of Gabor, in late 60's Willshaw started investigating the design of a distributed content-addressable memory on holographic principles [30]. In 1971 he proposed the correlograph model. However, it also used the simplified scalar representation instead of holographic multidimentional representation. This "correlograph" model is often referred to as "holographic". However, in technical sense it is closer to the hebbian learning based neural networks than the holographic memory discussed in Sutherland [26] Masters [20] and this paper.

3.4 Distinction of Holographic Computation

Transfer function: The above encoding and decoding algorithm can be realized in a distributed network of cells just like a conventional neuron, where each cell will be responsible for a simple computation with a transfer function of the form below:

$$\dot{z}_i = \sum_j^n \dot{w}_{ij}\dot{s}_j \ldots. \qquad (9)$$

Where all dot elements are MCN instead of scalars. The transformation it realizes on the measurement component of input information is fundamentally different from that of any existing AAM. Let, $\dot{w}_{ij} = \| w_{ij} \| e^{-\hat{i}\omega_{ij}}$. Then, the transformation between the measurement components of input and output is given by:

$$\phi_i = \cos^{-1}\left[\frac{1}{c}\sum_j^n \| w_{ij} \| \cos(\theta_j - \omega_{ij})\right]$$

$$\text{where, } c = \sum_j^n \| w_{ij} \| \qquad \ldots.(10)$$

For comparison, the scalar product rule of synaptic efficacy used by conventional AAMs is given below with equivalent notations:

$$\phi_i = f(y_i), \quad y_i = \sum_j^n w_{ij}\theta_j + b_i \qquad \ldots.(11)$$

This new transfer function has three characteristics that distinguishes itself from conventional transfer functions. The first is that the transfer function is a weighted *trigonometric (cosine) mean* function, in contrast to the conventional *weighted sum*. Secondly, that there is no explicit activation function. The third distinguishing feature of this cell is that all the individual synaptic inputs have their private thresholds, rather than having a single threshold at the output.

Synaptic transmission rule: The most important distinction is the first one. A finite summation process is tolerant to random statistical distortion, but is not tolerant to selective and deliberate loss of inputs. In contrast a mean process is robust in both the senses. This is the key distinction that allows a holographic cell, and thus a network of holographic cells, to conduct RCA search.

Mapping ability: The second and third distinctions are related and together these determine the mapping ability. In any associative memory the non-linearity decides the nature of the discriminating hyperplane that distinguishes classes. For a holographic cell the trigonometric transformation pairs serve as the implicit non-linearity. The only fundamental difference is that the non-linearity is local (like the existence of individual thresholds for each element) here. In contrast conventional neurons use global non-linearity which is applied after the weighted sum. Such localization of non-linearity is essential for attaining robustness against missing elements.

Hyperspherical representation: The fundamental distinction of holographic cell can be visualized from a representational perspective also. One of the basic limitation of the conventional network is that there is no representation of ’dont-care’. An element labeled as ’dont-care’ should be represented in such a way (state) that all the valid enumeration values of its measurements (states) should be equipotential. On a one dimensional (linear) space, it is not possible to obtain a point which is equidistant from all possible enumerations of an analog measurement. Any forced enumeration of ’dont-care’ on a real line will always induce undue bias towards two of the enumerations than all others. An obvious solution to this representation problem is to place the enumerations on a plane. MCN representation generalizes the above solution a step further and puts the enumerations on the surface of a hypersphere (Fig-3). The center enumerates an unbiased ’dont-care’.

The ability of this computational to perform the basic associative retrieval and also to satisfy the behavioral expectations of an attentive memory (outlined in section 2.1) has been formally shown in [10]. This paper now presents the performance analysis of it.

4. ANALYSIS OF PERFORMANCE

In this section the capacity and accuracy of this memory will now be measured.

Retrieval: The retrieved association can be decomposed into two parts; principal component and crosstalk component. This can be done by combining equations (6), (7) and (8).

$$[R^e] = \frac{1}{c}\cdot[S^e]\,[\bar{S}^t]^T\,[R^t] + \frac{1}{c}\cdot\sum_{\mu\neq t}^{P}[S^e]\,[\bar{S}^{\mu}]^T\,[R^{\mu}]$$

$$= [R^e_{principal}] + [R^e_{crosstalk}] \qquad \ldots.(12)$$

Where S^t is considered the candidate match (or target pattern). Both the principal and crosstalk components are derived below.

Principal Component: Individual elements of retrieved pattern are retrieved in identical manner independent of each other. Let us consider the retrieval of the u^{th} component of the response It is also assumed that all the encoded stimulus patterns have $\lambda = 1$.

$$r^e_{u(principal)} = \frac{1}{c}[S^e]\,[\overline{S}^t]^T r^t_j$$

$$= \frac{1}{c}\left[\lambda_1 e^{\left(\sum_j^{d-1} i_j \theta^e_{j,1}\right)}, \lambda_2 e^{\left(\sum_j^{d-1} i_j \theta^e_{j,2}\right)}, \ldots.\lambda_n e^{\left(\sum_j^{d-1} i_j \theta^e_{j,n}\right)}\right] \begin{bmatrix} 1.e^{\left(\sum_j^{d-1} \hat{i}_j \theta^t_{j,n}\right)} \\ 1.e^{\left(\sum_j^{d-1} \hat{i}_j \theta^t_{j,n}\right)} \\ . \\ . \\ . \\ . \\ 1.e^{\left(\sum_j^{d-1} \hat{i}_j \theta^t_{j,n}\right)} \end{bmatrix} r^t_j$$

$$= \frac{1}{c}\sum_k^n \lambda_k e^{\left(\sum_j^{d-1} \hat{i}_j\left(\theta^e_{j,k} - \theta^t_{j,k}\right)\right)} r^t_j \qquad \ldots(13)$$

If the query stimulus and the target stimulus corresponds closely, then for every *j* and *k* phase terms $\theta^t_{j,k} \rightarrow \theta^e_{j,k}$. Thus, all the exponent terms become unity with no phase disturbance. Which, reduces to,

$$r^e_{u(principal)} \cong \frac{1}{c}\sum_k^n \lambda_k r^t_u \qquad \ldots.(14)$$

The phase of the retrieved response corresponds to the retrieved information, and is equivalent to the phase of the encoded response:

$$argc(r^e_{u(principal)}) \cong argc(r^t_u) \qquad \ldots.(15)$$

.

Crosstalk Component: Similarly the crosstalk component is given by:

$$r^e_{u(crosstalk)} = \frac{1}{c}.\sum_{\mu \neq t}^p [S^e]\,[\overline{S}^\mu]^T\,[R^\mu]$$

$$= \frac{1}{c}\sum_{\mu \neq t}^p \sum_k^n \lambda_k e^{\left(\sum_j^{d-1} \hat{i}_j\left(\theta^e_{j,k} - \theta^\mu_{j,k}\right)\right)} r^\mu_u \qquad \ldots(16)$$

Saturation Ratio: The saturation ratio is defined as the ratio of the signal-to-noise magnitude[7]:

$$SR = \frac{|\, r^e_{u(principal)} \,|}{|\, r^e_{u(crosstalk)} \,|}$$

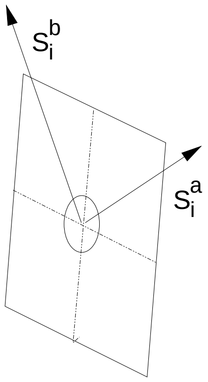


Fig-4 Angular Span of Elements

Let us consider that $P(a \times b, i)$ is the hyperplane spanned by the i^{th} elements of the a^{th} and b^{th} patterns (Fig-4). The orientation angle of element s^a_i in this plane will be denoted by $\psi^a_i\,|_{P(a \times b, i)}$. The difference between the orientation angles signifies the direct angular span between the elements s^a_i and s^b_i. Let us also define,

$$\phi^{e-\mu}_{k-l} = [\psi^e_k\,|_{P(e\times\mu,k)} - \psi^\mu_k\,|_{P(e\times\mu,k)}] - [\psi^e_l\,|_{P(e\times\mu,l)} - \psi^\mu_l\,|_{P(e\times\mu,l)}]$$

It denotes the difference between the angular spans between the k^{th} and l^{th} elements of the query(e^{th}) and μ^{th} stimulus patterns. It can be shown through some straightforward trigonometric manipulation that:

$$SR = \sqrt{\frac{\sum_k^n(\lambda_k)^2 + \sum_k^n \sum_{l \neq k}^n \lambda_k \lambda_l \cos\phi^{e-t}_{k-l}}{(p-1)\sum_k^n(\lambda_k)^2 + \sum_{\mu \neq t}^p \sum_k^n \sum_{l \neq k}^n \lambda_k \lambda_l \cos\phi^{e-\mu}_{k-l}}}$$

Assuming, independent identical and symmetrical distribution of α-suit (θ^μ_j), over all the element space of all the enfolded patterns:

$$E[\cos\phi^{e-\mu}_{k-l}] = 0$$

or for sufficiently large *pn*:

$$\left[\sum_{\mu \neq t}^p \sum_k^n \sum_{l \neq k}^n \lambda_k \lambda_l \cos\phi^{e-\mu}_{k-l}\right] \rightarrow 0$$

Thus,

$$SR = \sqrt{\frac{1}{(p-1)}.\left[1 + \frac{\sum_k^n \sum_{l \neq k}^n \lambda_k \lambda_l \cos\phi^{e-t}_{k-l}}{\sum_k^n(\lambda_k)^2}\right]}$$

7 Note, that saturation ratio is not same as the signal-to-noise ratio (SNR). SNR is the ratio of the signal-to-noise measurements.

Let us define a distance measure between two patterns d such that, α-suit elements of the stimulus S^e and S^t are bounded by the distance d over the entire set, such that $|\psi_j^e - \psi_j^t| \le d$, for all j which implies, $0 \le |\phi_{k-l}^{e-t}| \le 2d$

If the distance between the candidate and query is large ($d^{(e,t)} \gg 0$), then:

$$SR \approx \sqrt{\frac{1}{(p-1)}}$$

.

On the other hand, for close match, ($d^{(e,t)} \to 0$):

$$SR = \sqrt{\frac{1}{(p-1)} \cdot \left[1 + \frac{\sum_k^n \sum_{l \ne k}^n \lambda_k \lambda_l}{\sum_k^n (\lambda_k)^2} \right]}$$

$$\approx \sqrt{\frac{n}{(p-1)}} . w$$

Where, $\boldsymbol{w}$ is attention strength.

*Definition 6: **Attention Strength** w refers to relative strength of the attention distribution over the element space, and is defined by:*

$$w = \frac{\left[\sum_k^n (\lambda_k) \right]^2}{n . \sum_k^n (\lambda_k)^2} = \frac{[E\{\lambda\}]^2}{E\{\lambda^2\}} \qquad \text{....(17)}$$

The attention strength $\boldsymbol{w}$ intuitively refers to the 'porosity' of the window frame. It varies from 0 to 1 and depends on the distribution of λ in the query field. For type-U search all w=1. Thus, when for all $\lambda_j \to 1$ during encoding:

$$SR = \sqrt{\frac{n}{(p-1)}}$$

The above result can be summarized as:

Result 1: For the attentive memory specified with equations (6), (7), and (8) with n stimulus elements and p stored patterns, and an unequal distribution of attention specified by the vector $\Lambda^e = [\lambda_1, \lambda_2, \ldots, \lambda_n]$, the saturation is given by:

$$SR \approx \sqrt{\frac{n.w}{(p-1)}} \qquad \text{....(18)}$$

When (i) $p \gg 1$, and (ii) the elements are symmetrically distributed in phase space. Here **w** refers to the 'porosity' of the attention distribution.

<u>**Accuracy of Retrieval**</u>: Now the accuracy of retrieval will be derived. The resultant response is given by the sum of principal and crosstalk components. The case is investigated assuming perfect query meaning S^e resembles closely to one of the stored pattern. From equation (12) it can be seen that the capacity is limited by the accumulation of crosstalks from increasing number of patterns. Let the crosstalk component be given by $r_N e^{\hat{i}\theta_N}$, and the principal component be given by $r_S e^{\hat{i}\theta_S}$. Here the angles correspond to the direct angular span of the components in $P(r_{principal} \times r_{crosstalk}, u)$ hyperplane. Then the error in phase (which represents the measurement) of the resultant component is given by:

$$\phi_e = \tan^{-1}\left[\frac{r_N \sin(\theta_N - \theta_S)}{r_S + r_N \cos(\theta_N - \theta_S)} \right] \qquad \text{....(19)}$$

Fig-5 illustrated the addition in hyperspherical space. The phase deviation is maximum when $\Phi_e |_{\max} = (\theta_N - \theta_S) - 90°$. Thus, for saturation given by equation (18), the maximum phase error is:

$$|\Phi_{error}|_{\max} = \sin^{-1}\left(\sqrt{\frac{pw}{n}} \right) \qquad \text{....(20)}$$

When (i) $p \gg 1$, and (ii) the input elements are symmetrically distributed in phase space.

Result 2: For a MHAC specified with equations (6), (7), and (8) with n stimulus elements and p stored patterns the maximum distortion due to crosstalk is given by equation (20), when (i) $p \gg 1$, and (ii) the input elements are symmetrically distributed in phase space.

The above analysis shows that the focus can be effectively (almost linearly) compensated with higher n or lower p. This result is very significant. Because even for a fixed size problem, it is possible to design a network with exponentially higher effective stimulus length (n) by various techniques (such as higher order encoding). The above analysis provides us the clue to select a suitable n for a particular application.

Notably, the performance of this memory is dependent on the symmetry of the element distribution in the phase space. The performance of conventional neural networks is tied with the uniformity of distribution. Highly correlated elements in patterns destroys uniformity and consequently the performance of many ANNs. But uniform distribution is a special case of symmetrical distribution and is more restrictive. It is possible to obtain symmetrical distribution without uniformity. This is because unlike real interval (which enumerates elements of conventional ANNs) phase space is harmonic. As a result MHAC performance is less restrictively tied with correlated data set.

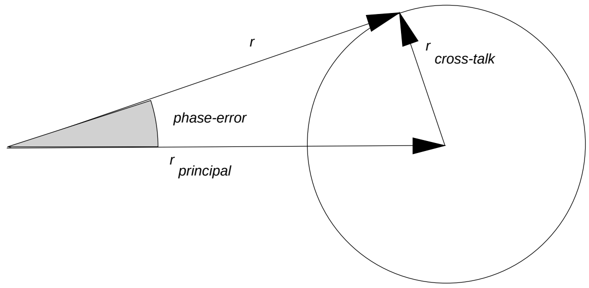


Fig-5 Geometry of Phase Error

Error due to imperfect query pattern: The two sources of error in the final response are (i) cross-talk due to saturation and (ii) principal component due to deviation of query pattern from the target pattern. Previous analysis showed the amount of error due to cross-talk. The error due to pattern deviation is the sum of the deviations of individual pattern elements. Thus it linearly moves away from the target patterns with mean of the shift in query $\Phi_\varepsilon = n.\bar{\varepsilon}_i$, when $\theta_i^t - \theta_i^e = \varepsilon_i$ when the error is small. It can be geometrically shown that the magnitude of the error due to pattern deviation grows in the order of $\sqrt{2\sin\left(\frac{\varepsilon}{2}\right)}$.

5. EXPERIMENTS

The analysis of last section shows that the performance of this new memory is dependent on (i) strength of focus, (ii) length of stimulus patterns, (iii) number of encoded patterns and (iv) distribution of data. This section presents a set of experiments to empirically validate and investigate the effect of each these factors. Below, first the parameters those have been used in these experiments are explained.

5.1 Parameters

Definition 7: ***Accuracy of Retrieval*** *(SNR) is measured as the peak signal to noise ratio in the measurement component of information over all the elements.*

$$SNR = 20\log\frac{2\pi}{mse} \qquad mse = \sqrt{\frac{1}{m}\sum_i^m[\phi_i^\mu - \phi_i^{T(\mu)}]^2}$$

The peak signal is given by the dynamic phase range 2π. *Average SNR* is computed by averaging over all the pattern associations enfolded in the memory.

Definition 8: ***Focus Strength*** *(f) is defined as ratio of the input significance strength of a query pattern S^e to the significance strength of encoded pattern.*

$$f = \frac{\sum_i^n \lambda_i}{n}$$

A uni-magnitude encoding of pattern elements has been assumed. Its value varies from 0 to 1. In the plots QPD=1-f has been used.

Definition 9: ***Load Factor*** *(L) is defined as the ratio of the total number of elements (n) in the patterns to the number of stimulus response associations (p) encoded.*

$$L = \frac{p}{n}$$

As evident, the length of stimulus (n) is already incorporated in the load factor.

Definition 10: ***Asymmetry*** *(k) of a pattern refers to the circular distribution of the pattern elements around the center of the representation hypersphere.*

$$k = \frac{\sum_i^n \lambda_i \exp\left(\sum_j^d \hat{i}_j \theta_i^j\right)}{\sum_i^n \lambda_i}$$

It is defined as above and its value varies from 0 to 1. In all these experiments, pattern elements have been generated randomly with *clipped Gaussian*[8] distribution to match natural distributions (such as image intensity). However, the standard deviation has been varied to generate data with different assymmetry characteristic. High standard deviation (SD) corresponds to low assymmetry and vice versa.

[8] Because of the circular nature of phase space, only those random generations have been used which falls between 0 to 2 Π.

Besides investigating the general relationship among these critical parameters, these experiments simultaneously examine the specific ranges of these parameters within which an effective and cost efficient attentive memory can be constructed.

5.2 ORS Experiment

The parameters of the attentive memory are independent, monotonic and together span a parameter space. The objective of the experimentation is to determine the sub-space (and their boundaries) of this parameter space within which it is possible to guarantee a target performance. It is called *operational range space* (ORS).

The availability of an ORS is advantageous from the engineering point of view. Given an ORS, when a new application is taken under consideration, all that is required, is to measure application specific parameters and to verify whether it falls inside or outside the ORS. If it is within, then the pre-analyzed results available from ORS experimentation can be used to predict approximate performance. Also the necessary configuration of the system for that application can be estimated. On the other hand, if it is outside, ORS experimentation results can still be used to identify the exact intervention that would bring the application within ORS.

5.3 Analysis of Experiments

Focus characteristic: The retrieval performance with the variation of focus strength is shown first. A set of holographs has been generated each with varied numbers of encoded patterns. After the training, by using a random part of each originally stored pattern as the query pattern, recalls have been performed. The focus strength has been controlled by varying the size of this part. For query pattern elements not selected in the focus set $\lambda_i \approx 0$ has been used. Fig-6 shows the typical average signal-to-noise ratio (left y-scale) and percentage of dynamic error (right y-scale) with the smooth variation of focus strength of the query pattern. The three curves in this graph show the focus characteristics for three different load factors L= .02, .04 and .08. For all these cases the patterns have length n=1000 and asymmetry SD=1.0 (k=.6). Fig-7 plots the performance for 3 different element distributions with respectively SD=.8,1.2, and 3 for L=.02 while other parameters remain same.

As evident from both of these plots, a typical focus characteristic curve is monotonic and resembles a fat sigmoid. These curves generally demonstrate three distinct zones, (a) high-performance zone, (b) linear-zone, and (c) cut-off zone.

The high performance zone corresponds to RCA type-U search performance. This zone characterizes regular AAM like high focus and is featured with high accuracy. As evident by the accuracy levels of this zone, an attentive memory, even when it acts as a regular memory far exceeds the retrieval accuracy of most other analog AAMs. This zone demonstrates accuracy over 30 dB (which in other words means less than 2-3% phase value error).

The most significant is the linear zone. In this zone, the accuracy gracefully decays with the focus strength. Analytically the characteristics of this zone correspond to equation (20). As can be seen, focus strength can be reduced almost as low as 0.1 till the accuracy falls below 20dB. In marked contrast, a regular AAM shows avalanche degeneration of performance when the focus strength approaches just 0.6-0.5 [27]. The cut-off zone for this attentive memory begins around 0.1 while that of conventional AAMs begins at 0.5. As can be observed in these plots, the typical ORS boundaries are (a) the high-performance zone extends from f=1.0-0.9, (b) linear-zone extends from f=0.9-0.1.

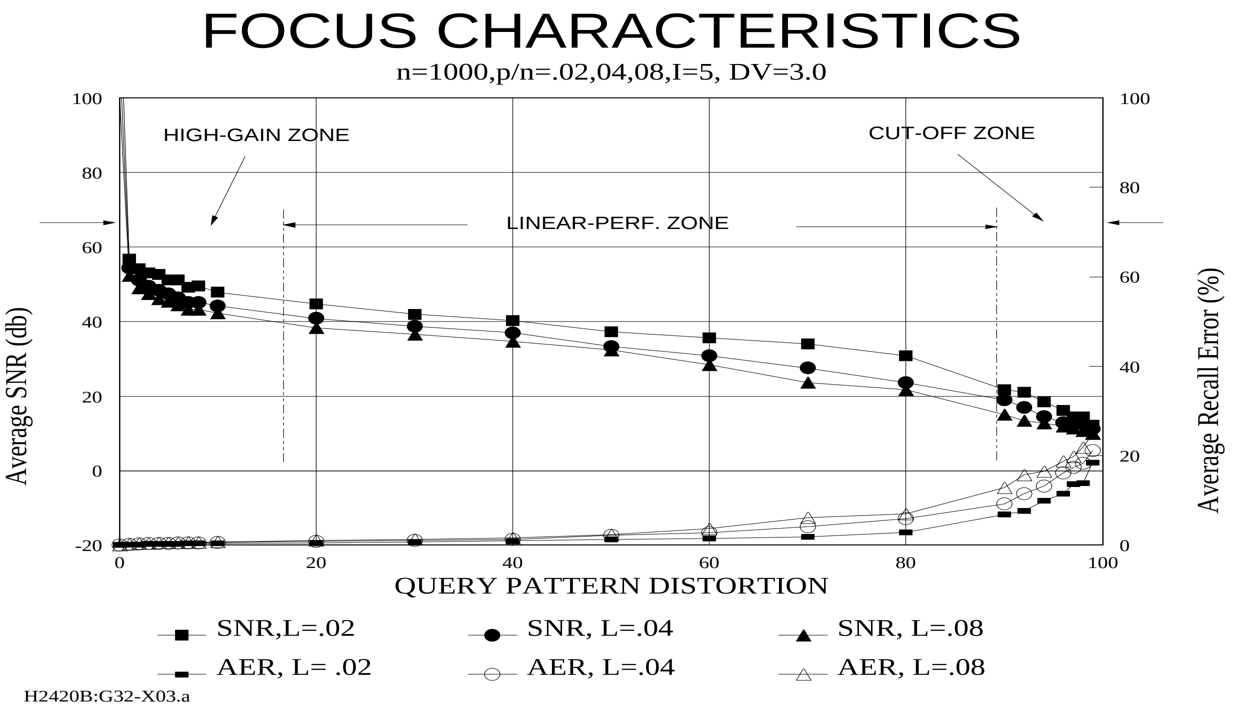

FOCUS CHARACTERISTICS
n=1000,p/n=.02,04,08,I=5, DV=3.0
HIGH-GAIN ZONE
CUT-OFF ZONE
LINEAR-PERF. ZONE
Average SNR (db)
Average Recall Error (%)
100
80
60
40
20
0
-20
0
20
40
60
80
100
QUERY PATTERN DISTORTION
SNR,L=.02
SNR, L=.04
SNR, L=.08
AER, L= .02
AER, L=.04
AER, L=.08
H2420B:G32-X03.a

Fig-6

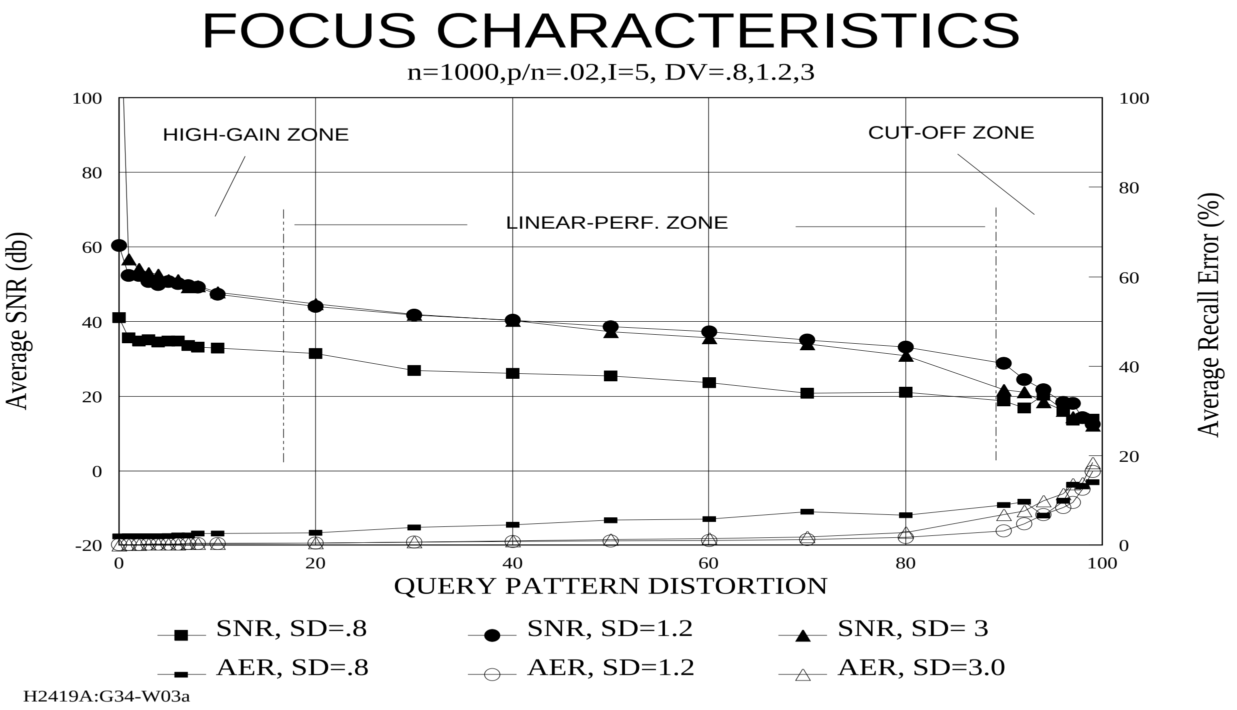

FOCUS CHARACTERISTICS
n=1000,p/n=.02,I=5, DV=.8,1.2,3
HIGH-GAIN ZONE
CUT-OFF ZONE
LINEAR-PERF. ZONE
Average SNR (db)
Average Recall Error (%)
100
80
60
40
20
0
-20
0
20
40
60
80
100
QUERY PATTERN DISTORTION
SNR, SD=.8
SNR, SD=1.2
SNR, SD= 3
AER, SD=.8
AER, SD=1.2
AER, SD=3.0
H2419A:G34-W03a

Fig-7

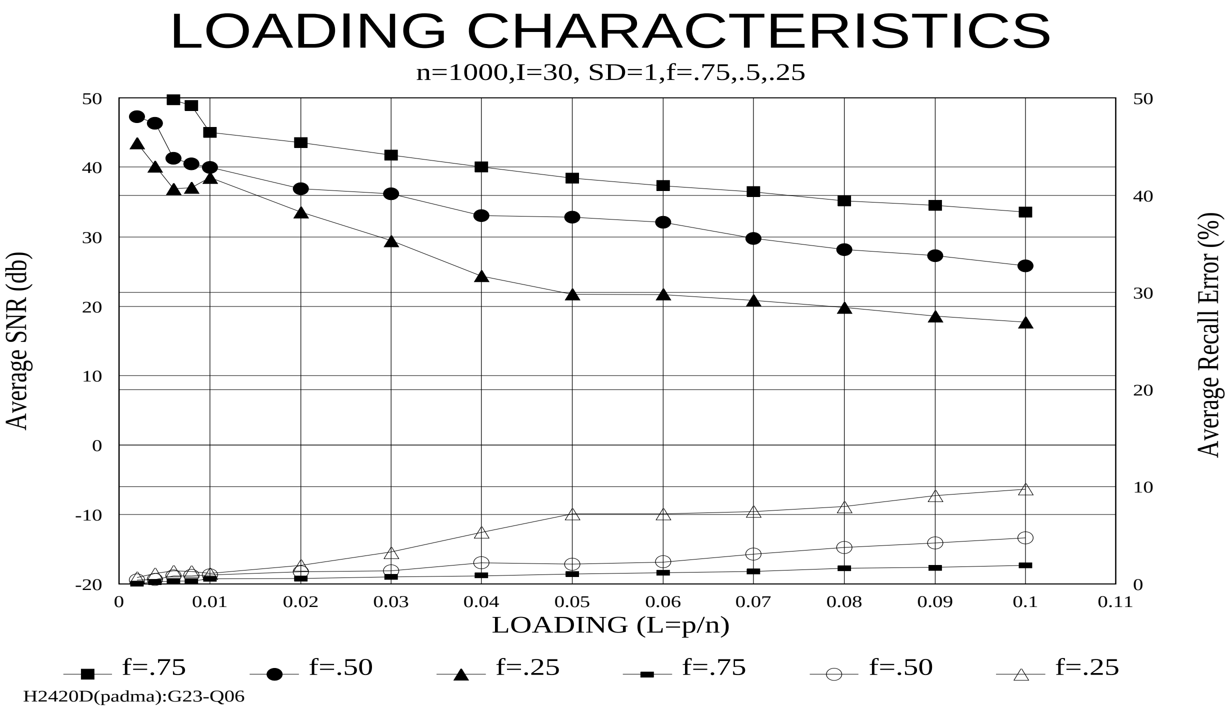


Fig-8

Loading characteristic: Next simulation shows the effect of loading on the performance of the attentive memory. To determine the ORS boundaries of the load factor range, first a pool of clipped Gaussian patterns has been generated (all with a fixed length). Then different holographs have been trained each time taking a different number of patterns from this pool.

Fig-8 shows a usual loading performance. It plots the SNR (y-axis) against various load factors (x-axis) for three RCA type-A cases with focus strengths f=0.75, f=0.50, and f=0.25. The pattern sets are generated with standard deviation SD=1.0. The average asymmetry (k) of these patterns is found to be k=0.3.

As shown in this plot, a typical loading characteristic curve shows monotonically decreasing performance with increased load factor. Quantitatively, for f=0.25, the RCA type-A retrieval accuracy drops to 20 dB, while the load factor reaches 0.07. Typically a load factor of 0.03 to 0.10 can be reached maintaining 20 dB performance with f=0.3-0.1. This range defines the load factor boundary of ORS. This experiment shows that an enormous number of pattern associations can be stored and retrieved from a single holographic memory. For example, a load factor of 0.02 means that about 5,000 images of size 512x512 can be enfolded into a single holographic attentive memory and can be searched with RCA type-A capability. This particular search (which is equivalent to searching into 1.25 GB of raw space) can be done only at the cost of only one complex inner product[9]. Table-4 lists few other loading scenarios.

9 The size of the matrix is nxm, where m is the response label size and typically is $O(\log p)$, where as a procedural best match search is nxp.

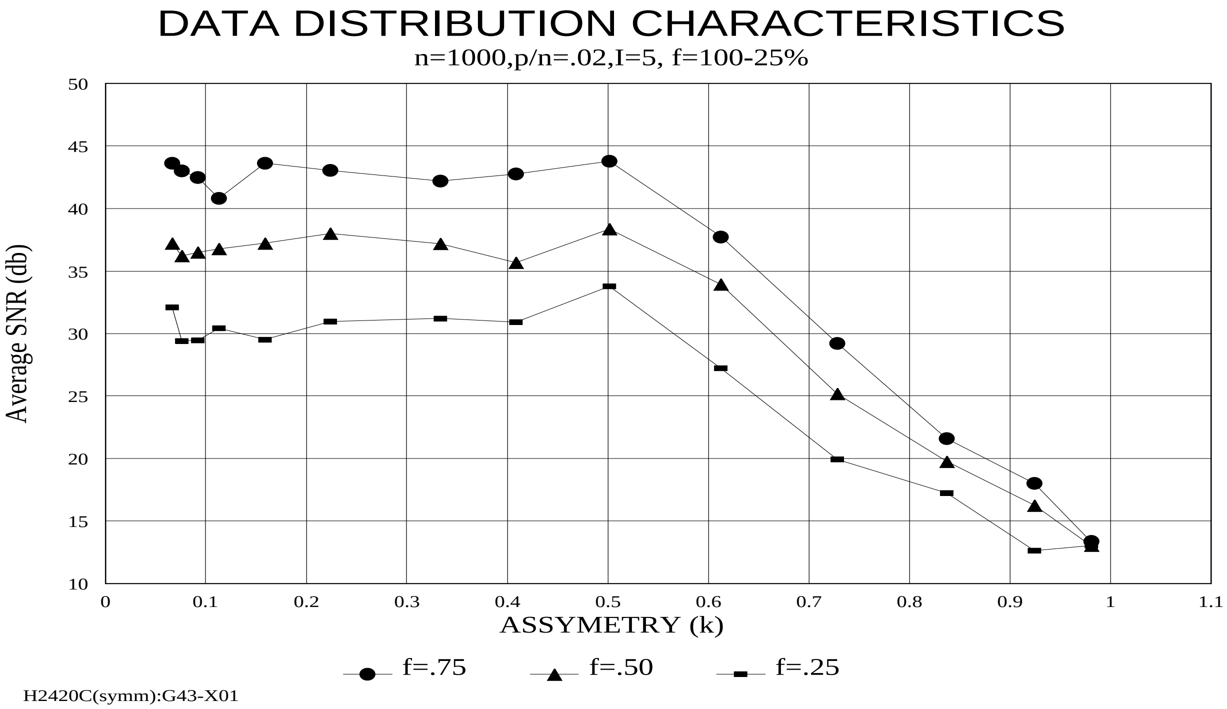


Fig-9

Asymmetry characteristics: The ORS boundaries of asymmetry parameter can be observed through the projection of range-space by continuous variation of k. To perform this experiment, several sets of patterns have been generated with varying standard deviations. These are then encoded into different holographs. Variation in the standard deviation (of clipped Gaussian distribution) generates data sets with various asymmetries. The narrower the deviation, the higher the asymmetry.

Fig-9 shows a typical data distribution characteristic. It plots the SNR (y-axis) against the primary parameter asymmetry (x-axis). Three RCA type-A test results with secondary parameter focus strength f=0.75, f=0.50, and f=0.25 have been shown by the three curves. For these experiments, each of these holographs has been loaded with a load factor =0.02, and has been trained with 5 iterations.

Typically, as the asymmetry increases, the performance of MHAC decreases. As shown in Fig-9, MHAC can tolerate up to 0.6 asymmetry of the data distribution and can still maintain 20 dB performance within the operational range-space.

The result of this experiment is particularly important for the design of the mapping transform. The actual nature of data distribution depends on the application and in most cases is beyond the control of the system designer. Table-3 shows some examples of asymmetry values for few typical images. In the extreme cases of unusually ill skewed data set, the above result provides important guideline to the designer. Appropriate transforms $m^{+}(.)$ can be designed by which the asymmetry level of the processed data can be hashed within the acceptable range.

Effect of stimulus length: As evident from the definition, the stimulus length is already a part of load factor. Therefore, the principal effect of *n* can be observed readily in Fig-8. However, an important remaining question is whether the performance of holographic attentive memory is sustainable for larger scales of this memory with larger values of p and n even when the ratio is fixed. This experiment is particularly designed to investigate such scalability of attentive memory.

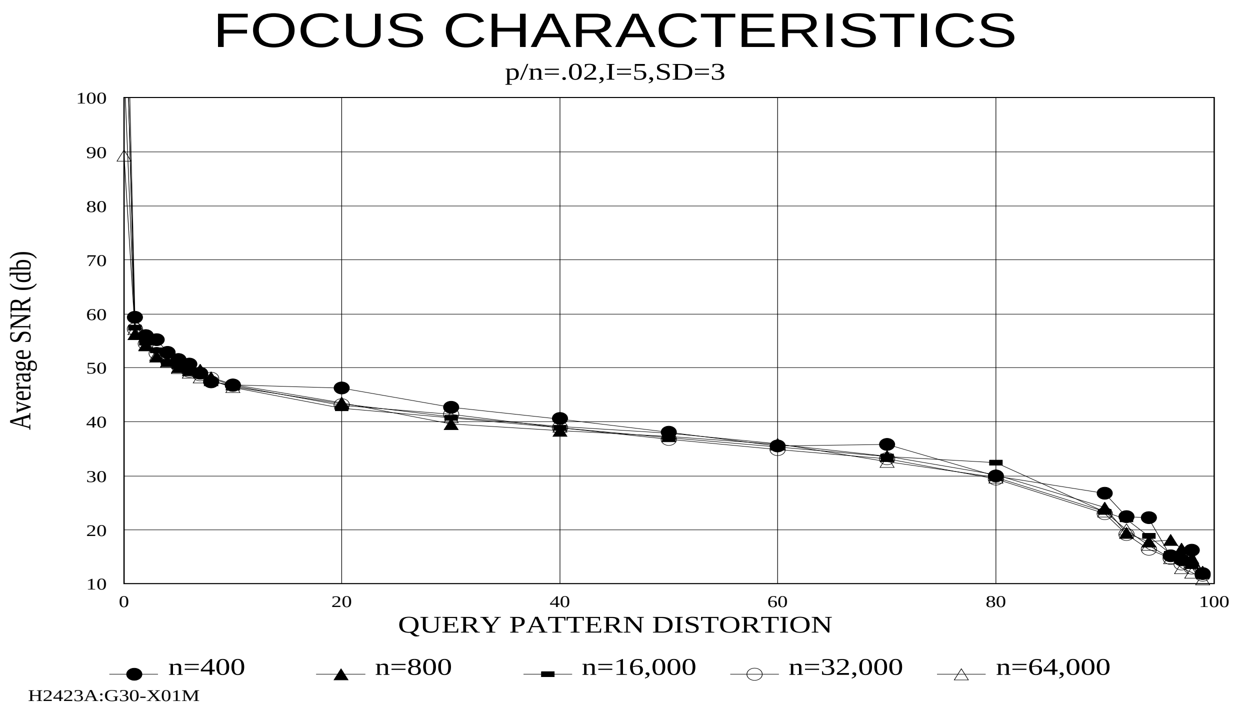

FOCUS CHARACTERISTICS
p/n=.02,I=5,SD=3
Average SNR (db)
100
90
80
70
60
50
40
30
20
10
0
20
40
60
80
100
QUERY PATTERN DISTORTION
n=400
n=800
n=16,000
n=32,000
n=64,000
H2423A:G30-X01M

Fig-10

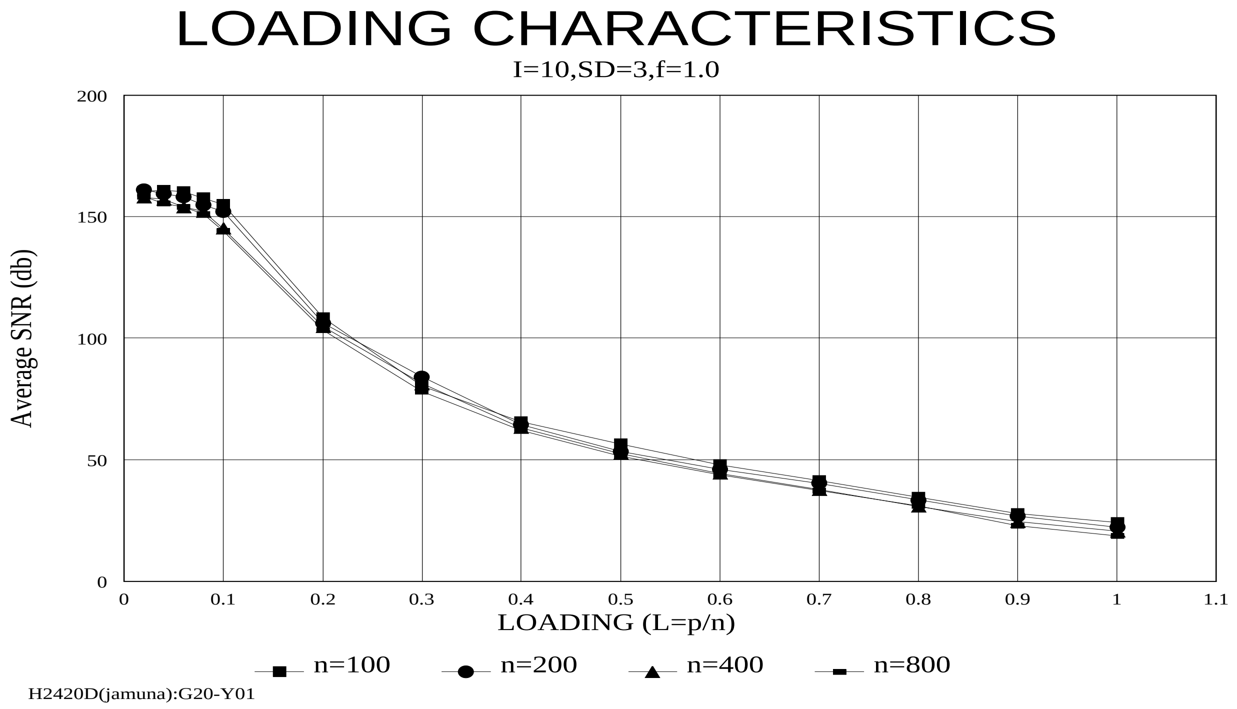

LOADING CHARACTERISTICS
I=10,SD=3,f=1.0
Average SNR (db)
200
150
100
50
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
LOADING (L=p/n)
n=100
n=200
n=400
n=800
H2420D(jamuna):G20-Y01

Fig-11

Fig-10 plots the SNR against focus strength *f* for exponentially varying *n*=400, 800, 1600, 3200 and 6400 for a fixed *p/n* ratio. As evident, although the problem scale varies exponentially, these curves overlap on each other demonstrating the performance invariance of focus characteristics with respect to the scale of *n*. Similarly Fig-11 shows the scale invariancy of loading characteristics. The result has been verified by repeated simulation for all the other plots for other scales.

This result also conforms to the analytical derivations (equation (18), and (20) which show that most of the performance characteristics are related to the load ratio *p/n*, rather than *n*. Both analytical as well as experimental results indicate the enormous algorithmic scalability of the attentive memory with sustainable performance.

Dimensionality of representation: The objective of this final set of simulation is to examine the effect of representation dimension (the dimension of hyperspherical space in which the element vectors are oriented) on the performance of the attentive memory.

For this experiment the same sets of randomly generated vectors have been encoded with different dimensionality of representation. To generate asymmetry variations, the orientation phases of these vectors are mapped uniformly within $+R \geq \theta > -R$ range during mapping. Five specific distributions with $R = 45^0, 60^0, 90^0, 120^0, 180^0$, have been considered. Narrower R corresponds to narrower distribution range and thus higher asymmetry. The experiment has been performed with P (=16) vectors each with D dimensional S (=32) elements. Each of the phase components has been generated randomly with uniform step distribution within the range R. During the recall process the principal component and the cross component of the separately measured. The experiment is repeated for dimensions D=2 to 10.

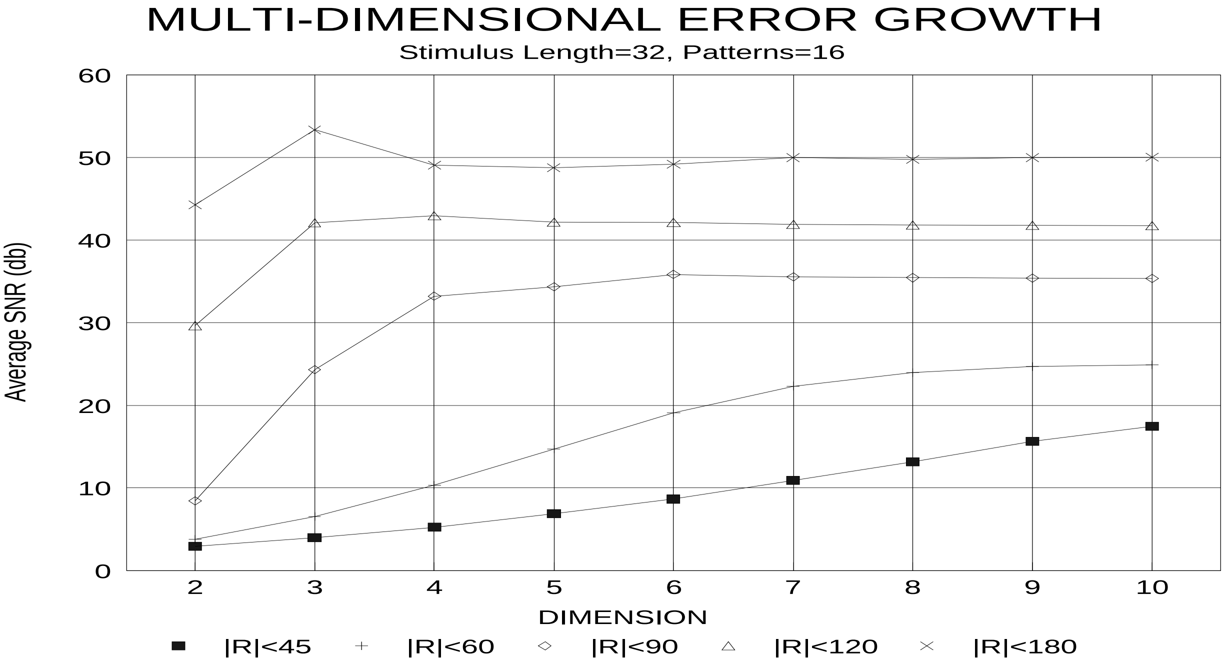


Fig-12 SNR and Dimension

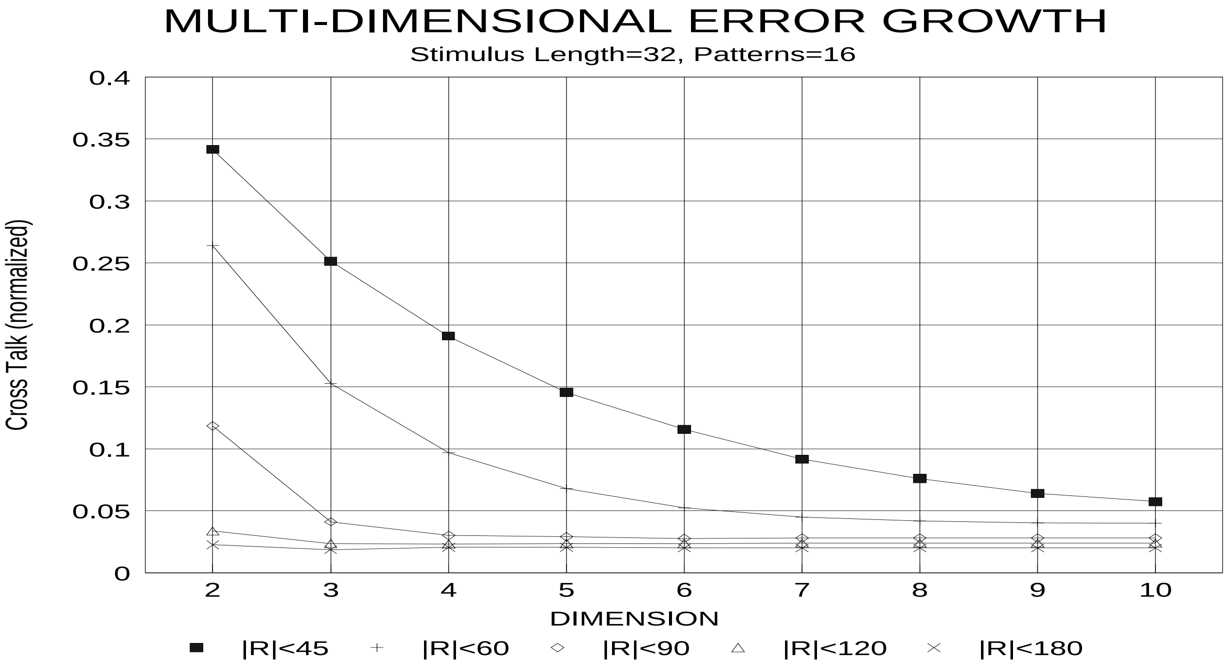


Fig-13 Crosstalk and Dimension

Fig-12 plots the signal to noise ratio against the dimensionality. Fig-13 plots the crosstalk component against dimensionality. The results of this experiment clearly show that SNR improves (Fig-12) and crosstalk (Fig-13) decreases as one shifts to higher dimensional representation. It is also evident that the improvement is more drastic if the phase distribution window is narrow.

This experiment helps in appreciating the contribution of multidimensional representation over associative computing. A 2D representation space helps in incorporating the novel notion of attention into associative memory. Thus it makes a qualitative difference over the capabilities of representationally scalar associative computing. Higher dimensions can further increase it's performance[10], and thus makes additional quantitative improvement.

5.4 Summary of ORS Boundaries

The quantitative results of ORS experiments are summarized in Table-2 suggesting an operating range space (ORS) for the 2D attentive memory system. Table-2 in particular guarantees an accuracy in the range of 20 dB. It shows the asymmetry, load factor, focus and iteration ranges required to achieve this target performance. All these parameters are monotonic, hence the space spanned by these boundary values and the co-ordinate planes represents the ORS.

Parameter	Unit	Operational range
Accuracy	SNR(dB)	>20
Asymmetry	*k*	<.6
Load factor	*L*	<.08
Focus	*f*	>0.1

Table-2 Operational Range Space

In Table-2, accuracy is the target parameter. Asymmetry is a data dependent parameter and is a semi-controllable constraint in a given application. It is possible to improve symmetry through various smoothing techniques. Table-3 shows some asymmetry measure of few well-known[11] example images.

image	*k* (1st order)	dimension
lake	.23	512x512
tree	.19	256x256
lena	.38	256x256
house	.27	256x256
pepper	.17	512x512

Table-3 Typical Asymmetry

[10] This characteristic has been analytically validated in details in [10].

[11] Can be downloaded from author's home page.

Loading and training are two controllable design parameters. Loading is closely tied to the space efficiency of any associative memory. The dimension of a holograph is determined by the length of the stimulus (*n*) and response patterns (*m*). Load factor provides an estimate how many such patterns can be enfolded on a single holographic memory. Table-4 shows typical estimates on the number of patterns that can be stored (and queried) for few image sizes. However, for patterns with limited size, load factor is not necessarily a hard limitation. The number of stored patterns *p* for relatively small patterns can be increased by higher order encoding.

n	*L*	*p* (1st order)
160×120	.04	768
256×256	.02	1310
512×512	.02	5120
1024×1024	.01	10240
1024×1024	.02	20480

Table-4 Typical Memory Loading

The above operational range-space provides a quick means for predicting the performance and estimating design parameters whenever a new application is considered. For example, if an associative memory with CT-scan images is to be constructed, all that is required is to estimate the asymmetry characteristics of the images. If the asymmetry is within the range space (k<0.6), then it is possible to predict the required dimension and other parameters for the corresponding attentive memory system. On the other hand, if k>0.6, even then it is possible to estimate how much smoothing is needed to obtain the target performance.

5.5 An Associative Search Example

The attention ability of this model is now demonstrated through an image pattern retrieval example. An MHAC memory has been created with 64 CT scan and MRI images each of size 256x256 pixels. Fig-15 shows the full frame retrieval accuracy of this memory for each of the 64 stored images. Fig-14(a) shows a typical query image with two windows, each focusing on a cognitively significant object in it (*Vertebrea* and *Kindney*). Table-5 lists the visual specifications of these objects in terms of their four corners and their size relative to the frame. Fig-14(b) shows the corresponding matching images which has been respectively retrieved by the memory as the associative match. As evident, based on the focus specification, each time the memory correctly retrieved the appropriate target image. Although none of these stored pictures have global statistical similarity with the query image, but both the matches are correct on the basis of localized similarity. Table-6 shows the performance. As evident in these cases, often the focus strength of effective cue lies in the range of 5-20% of the entire frame size and MHAC can retrieve them with 20-40 dB accuracy.

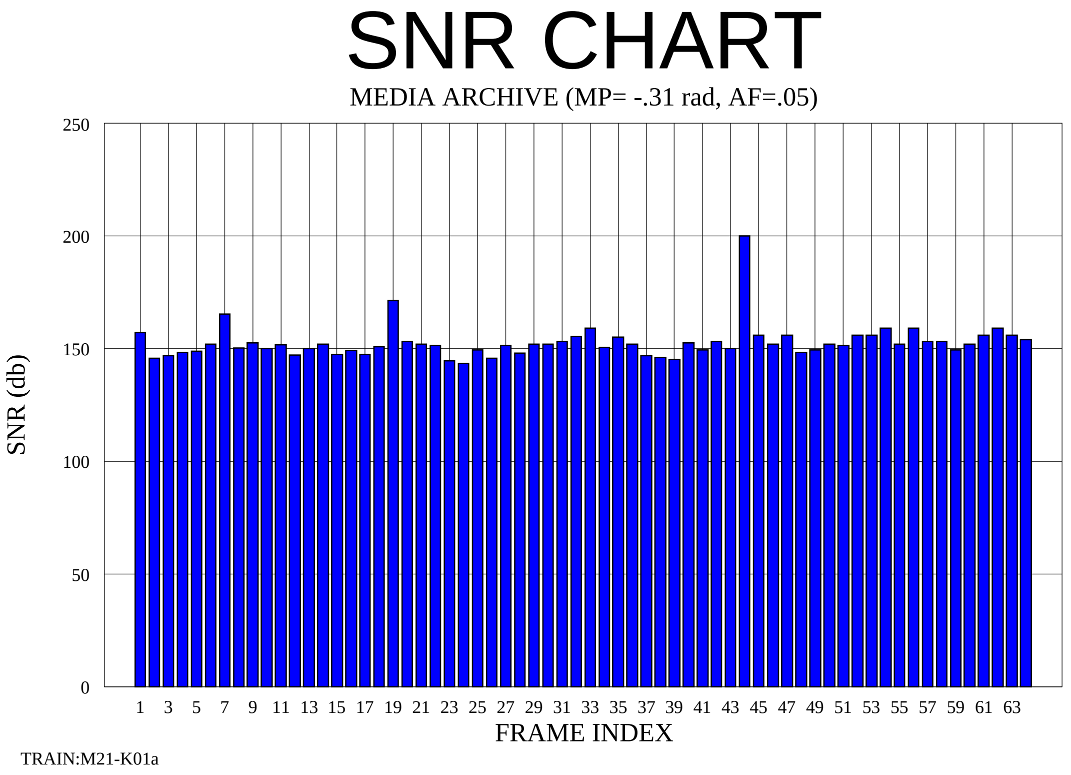


Fig-15 Decoding Accuracy

MASK#	Object	xmax	xmin	ymax	ymin	rho
4	Kidney	230	-135	165	-067	.145
6	Vertebrae	119	-045	145	-031	.131

Table-5 Specifications of the Windows of Focus

MASK#	Object	f	Match#	SNR (db)
4	Kidney	.145	4.4 (A26)	31.22
6	Vertebrae	.131	6.1 (A25)	27.55

Table-6 Results of Retrieval

Most conventional AAMs are currently unable to support the demonstrated retrieval for two reasons. First, the cue sizes are far below the 50% statistical dominance barrier of current AAMs. Secondly, and more profoundly, they would not be able to converge to diverse matches from same input pattern since they do not support dynamic search localization.

6. CONCLUSIONS

The paper investigates the performance characteristics of a new associative memory called attentive memory both analytically and through computer simulations. The results strongly suggest both quantitative and qualitative advantages of this new memory over existing parallel and distributed models of associative computing. In this concluding section the principal results and their implications will be briefly summarized.

Performance: It has been shown quantitatively that as low as 5-10% cue can be effectively used. This is a fundamental improvement over the capabilities of existing AAMs. Existing AAMs cease to retrieve when the valid part of the query pattern falls below the statistical limit of 70-50% the original pattern size [27,8]. The ORS ranges for focus (f=1-.1) and loading characteristics (L=.01-.04) suggests the designability of real (associative memory based) applications with this model. It has also been shown that its performance as a regular type-U exceeds that of most of the existing type-U equivalent models [26,10].

Scalability: Any pseudo optimization algorithm, besides the sustenance of speedup (architectural scalability), also requires sustenance of the quality of the solution for scalability (computational scalability). Analytical and empirical evidences obtained in this work both suggest that the performance of the network is sustainable for larger scale of the problem size (characterized by *n* and *p*). It is a well-observed phenomenon that the scalability of even the most successful ANN models (such as Backpropagation, Counterpropagation networks) are rather limited. Not only the amount of computation increases, but also the convergence speed, accuracy of any conventional ANN degrades steeply when problem size increases. In addition to the demonstrated computational scalability, the highly structured and heavy grain complex valued matrix operations of this memory makes it suitable for parallelization[12] and suggests simultaneous architectural scalability.

Implementation: As evident, the computational model of this memory is highly structured and repetative. Such characteristics make this entire model implementable with easily cascadable and reusable VLSI blocks. We are currently investigating an integrated architecture using hierarchical shared memory with a set of concurrent encoding/decoding processors. As already explained, at the macro level these same properties favor highly parallel and distributed implementation on conventional MIMD parallel machines. This memory also bears excellent potential for optical realization. The hyperspherical computations maps naturally on optical computations. Also, in the optical realization, patterns can be recalled by non-mechanical means. Which signifies that the access time can be in the order of several microseconds (which is 100-1000 times faster than current Compact Disk technology). Recently optical holographic echnology has made phenomenal advancement as storage medium [23]. As evident, the results obtained in this work directly broaden the computational potential of this promising (and ripe) technology by demonstrating their more advanced applicability in dynamic attention based associative recollection.

Potential Applications: Qualitatively this new memory provides the novel RCA type-A and type-B capability within the framework of associative computing. It can potentially facilitate the solution of a whole new class of unresolved problems requiring both adaptability of model acquisition and dynamic associative recollection[13]. *Content based retrieval in image archive, search in massive digital libraries, target recognition, pattern analysis in multidimensional spectral data, associative inference engine, real time speech synthesis* [12,3] are just few of the daunting problems which fit in this class and can directly benefit from this new memory model with attention. MHAC has already been successfully used to develop the first associative memory based approach for a content-based image archival and retrieval system. This approach can overcome subjective incoherence of traditional symbolic model mediated approaches [12,13,14].

The success of any computational model as a knowledge hub will require much more flexible and sophisticated retrieval capabilities than those which are offered by today's neural computing, in addition to learning and knowledge acquisition capabilities. Current

[12] Current generation parallel computers are characterized by their regular and structured architecture and relatively high communication cost. As a result they favors computations which are regular and heavy grain.

[13] If we analyze the successful applications of existing AAMs that evolved over last two decades, it will be evident that most of these use AAMs as adaptive filters and classifiers. Cosequently, in current literature neural networks are often referred by the term 'adaptive filter' almost as a synonym [1]. However, hardly any successful application exists which truly utilizes the associative memory property of AAMs. Such state of the art indeed reflects, on one hand, the sophistication of the learning ability, on the other hand, the constraints of the associative recollection ability of current AAMs.

models excel mostly in the later. The demonstrated attentive memory takes current associative computing a step closer to that goal.

The research has been supported by an East West Center Fellowship. Also a part of this research has been funded by ACTS and Supercomputing in Remote, Cooperative Medical Triage Support and Radiation Treatment Planning project of ARPA under research grant DABT 63-93-C-0056.

7. REFERENCES

[1] Carpenter, G. A., "Neural Network Models for Pattern Recognition and Associative Memory", *Neural Networks*, v.2, 1989.

[2] Carpenter G. A., S. Grossberg, N. Markuzon, J. H. Reynolds, & D. B. Rosen, "Attentive Supervised learning and Recognition by Adaptive Resonance Systems", *Neural Networks for Vision and Image Processing*, Ed. G. A. Carpenter, S. Grossberg, MIT Press, 1992, pp364-383.

[3] Chang, S.K., Arding Hsu, "Image Information Systems: Where Do We Go From Here?", *IEEE Trans. on Knowledge and Data Engineering*, v.4, n.5, October 1992, pp431.

[4] Gabor, D., "A New Microscopic Principle", *Nature*, v.161, 1948, pp777-778.

[5] Gabor, D., "Associative Holographic Memories", *IBM Journal of Research and Development,* I3, 1969, pp156-159.

[6] Grossberg, S., "Nonlinear Difference-Differential Equations in Prediction and Learning Theory", *Proc. of National Academy of Science*, v.58, n.4, October 1967, pp1329-1334.

[7] Grossberg, S., "On the Development of Feature Detectors in the Visual Cortex with Applications to Learning and Reaction-Diffusion Systems", *Biological Cybernetics,* v.21, n.3, 1976, pp145-159.

[8] Hopfield, J. J., "Neural Networks and Systems with Emergent Collective Computational Abilities", *Proc. of National Academy of Science*, USA, v.79, April 1982, pp2554-2558.

[9] Jacobson, M, *Foundations of Neuroscience,* Plenum Press, New York, 1993, pp173.

[10] Khan, Javed. I., "Attention Modulated Associative Computing and Content Associative Search in Images", *Ph.D. Dissertation,* Department of Electrical Enginnering, University of Hawaii, July, 1995.

[11] Khan Javed. I., and D. Y. Y. Yun, "Chaotic Vectors and a Proposal for Multidimensional Complex Associative Network", *Proceedings of SPIE/IS&T Symposium on Electronic Imaging Science & Technology '94, Conference 2185,* San Jose, CA, February 1994, pp95-106.

[12] Khan Javed. I.,& D. Yun, "Searching into Amorphous Information Archive", *International Conference on Neural Information Processing*, ICONIP'94, Seoul, October, 1994, pp739-749.

[13] Khan J. I.,& D. Yun, "An Associative Memory Model for searching Image Database by Image Snippet", *Proceedings SPIE Conference on Visual Communication*, VisCom'94, Chicago, September, 1994, pp591-601.

[14] Khan J. I.,& D. Yun, "Feature Based Visual Query in Image Archive with Holographic Network", *Proceedings of the International Conf. on Robotics, Control and Vision*, ICARCV'94, Singapore, November, 1994.

[15] Klopf, A. H., "Drive-Reinforcement Learning: A Real Time Learning Mechanism for Unsupervised Learning", *Proc. of 1st IEEE Conf. on Neural Networks,* Vol.II, N.J., 1987, pp441-445.

[16] Kohonen, T., Content Addressable Memories, 2nd Ed., Springler Verlag, Berlin, 1987.

[17] Kohonen, T., *Self-Organization and Associative Memory,* 3rd Ed., Springler Verlag, Berlin, 1989.

[18] Kumar, B. V. K., P. H. Wong, "Optical Associative Memories", *Artificial Neural Networks and Statistical Pattern Recognition,* I. K. Sethi and A.K. Jain (Eds.), Elsevier Science Publisheres, 1991, pp219-241.

[19] Leigth, E. N., and J. Upatnieks, "Photography by Laser", *Sceintific Amearican*, June 1965.

[20] Masters, T., *Signal and Image Processing with Neural Networks*, John Wiley & Sons, New York, 1994.

[21] McCulloch, W. S., Walter H. Pitts, "A Logical Calculus of the ideas Immanent in Nervous Activity", *Bulletin of Mathematical Biophysics* v.5, 1943, pp115-133.

[22] Oja, E., "A Simplified Neuron Model as a principal Component Analyzer", *Journal of Mathematical Biology,* v.15, 1982, pp267-273.

[23] Psaltis, D, Fai Mok, Holographic Memories, *Scientific American*, November 1995, pp70-76.

[24] Whitehead, A. N. and B. Russell, *Principia Mathematica,* 2d ed. Cambridge, Cambridge University press, 1927.

[25] Sherrington, C.S., *The Integrative Actions of Nervous System*, Yale Univ. Press, New Haven, 1906.

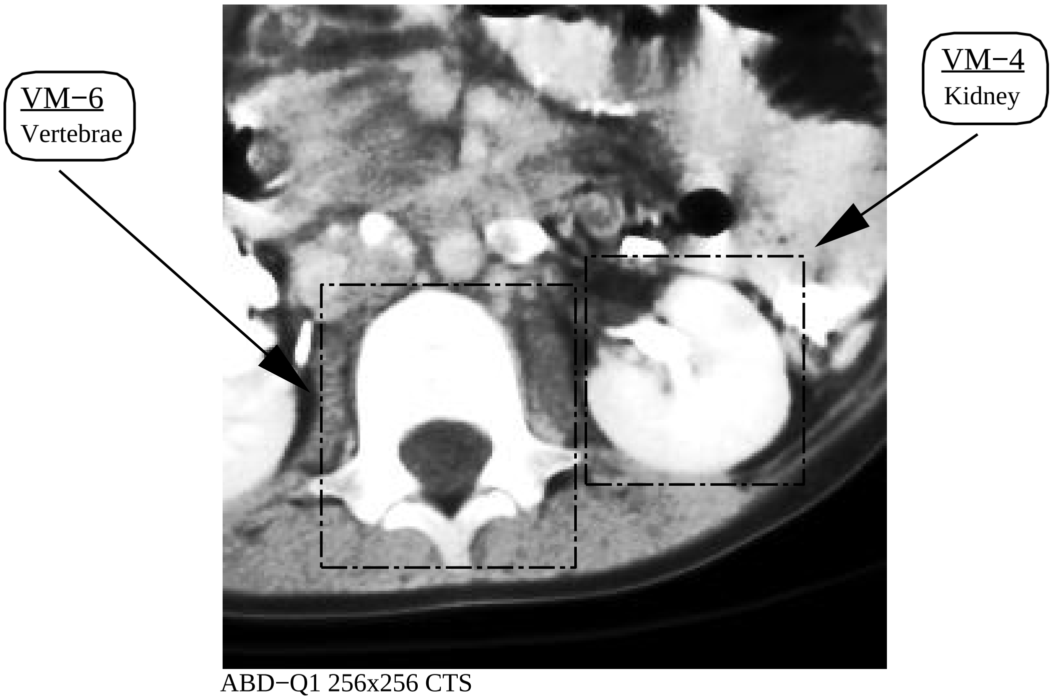
[26] Sutherland, J., "Holographic Models of Memory, learning and Expression", *International J. Of Neural Systems*, 1(3), 1990, pp356-267.

[27] Tai, Heng-Ming, T. L., Jong, "Information Storage in High-order Neural Networks with Unequal Neural Activity", *J. of Franklin Institute,* v.327, n.1, 1990, pp16-32.

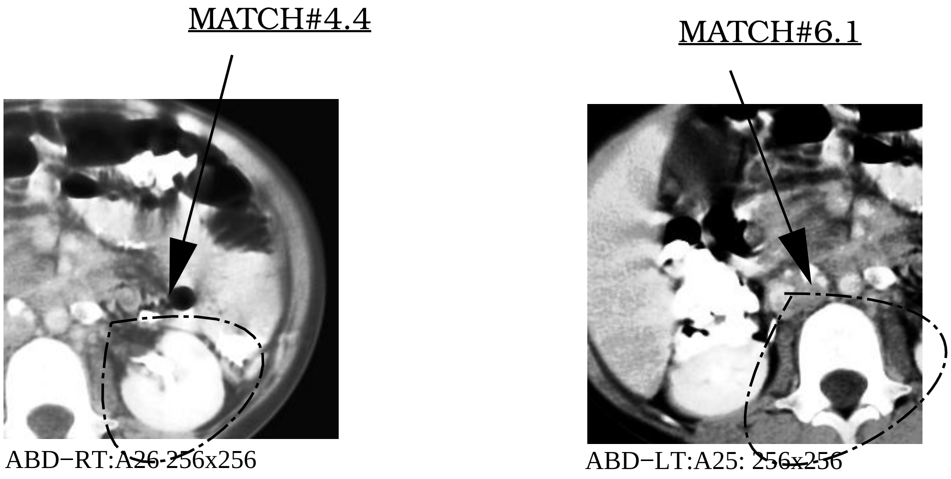
[28] Wenyon, Michael, Understanding Holography, Arco Publishing Inc., NY 1978.

[29] Widrow, B., M.E. Hoff, "Adaptive Switching Circuits", *IRE WESCON Convention Record,* part 4, 1960, pp96-104.

[30] Willshaw, D., Holography, associative memory and inductive generalization, in Parallel Models of Associative Memory, G.E. Hinton and J. E. Anderson, Eds, Hillsdale, NJ: Erlbaum, 1985.



Fig−14(a) Sample Query Image and Focus Objects



Fig−14(b) Retrieved Images from the Attentive Memory