Regular Paper

NI can do better compressive sensing

Harold Szu

Catholic University of America
szuharoldh@gmail.com

Abstract

Based on Natural Intelligence (NI) knowledge, our goal is to improve smartphone imaging and communication capabilities to resemble human sensory systems. We propose adding an enhanced night imaging capability on the same focal plane array (FPA), thus extending the spectral range. Compressive sensing (CS) technology reduces exposure to medical imaging and helps spot a face in a social network. Since Candes, Romberg, Tao, and Donoho (CRT&D) publications in 2007, 300 more contributions have been published in IEEE. What NI does differently is mimic the human visual system (HVS) in both its day-night and selective attention communication capabilities. We consider two killer apps exemplars: Software: generating video Cliff Notes; Hardware: designing day-night spectral camera. NI can do better CS, because of connectionist build-in attributes: fault tolerance filling missing parts; subspace generalization discovering new; unsupervised learning improving itself iteratively.

Keywords: Unsupervised Learning, Compressive Sensing, HVS, Smartphone, Fault Tolerance, Subspace Generalization, Medical Image, Face Identification.

1. Introduction to Compressive Sensing and Natural Intelligence Technologies

Compressive Sensing: Compressive Sensing (CS) technology is motivated by reducing the unneeded exposure of medical imaging, and finding a parse representation to spot a face in social nets. A large community of CS has formed in the last 5 years, working actively on different applications and implementation technologies. The experience of the International Neural Network Society (INNS) working on traditional computational systems in the last decades developing the unique capabilities of unsupervised learning, cf. Sect. 2, can be beneficial to a larger community, if a concise treatise of learning is made available. Likewise, INNS can benefit from the mathematical insights and engineering implementations of CS.

Face Spotting App: To find a friend, one may turn on a smartphone app for spotting a friendly face among the crowd, or simply surf in Facebook. Such a spotting app may be built upon a massive parallel ANN System On Chip for face detection (SOC-FD), which detects all faces (by color hue pre-processing) in real time and simultaneously places all faces in boxes in 0.04 seconds and identifies whom is smiling and who is not and closed eyes, focusing on the smiling one (by the fuzzy logic post-processing). Each high resolution image on a smartphone has mega pixels on target (pot). Each face has a smaller pot denoted by $N \cong 10^4$. Since the FD-SOC can cut equal-size facial pictures $\{\vec{x}_t, t = 1,2,3,4\ \}$ at different poses, likewise, the other person $\{\vec{y}_t, t = 1,2,3,4\}$, etc., if so wishes, forms a database $[A]_{N,m} = [\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4, \vec{y}_{1,}., \vec{z}_1, .]_{N,m}$ with m= *3x4* faces. The app CS algorithm matches an incoming face $\vec{Y}_N$ with the closest face in the database $[A]_{N,m}$. Mathematically speaking, since the database $[A]_{N,m}$ is over-sampled, finding a match is equivalent to finding a sparse representation $\vec{X}_m$ of $\vec{Y}_N$, i.e. $\vec{X}_m$ has few ones (=yes) matched, among many mismatched zeros (=no).

$$\vec{Y}_N = [A]_{N,m}\vec{X}_m \qquad (1)$$

Yi Ma *et al*. of UIUC have further applied a down-sampling sparse matrix $[\Phi]_{m,N}$ to the database $[A]_{N,m}$ for the linear dimensionality reduction for real-time ID in IEEE/PAMI 2007.

Medical Imaging App: Emmanuel Candes of Stanford (formerly at Caltech), Justin Romberg of GIT, and Terrence Tao of UCLA [1,2] as well as David Donoho of Stanford [3] (CRT&D) jointly introduced the Compressive Sensing (CS) sparseness theorem in 2007 IEEE/IT, in order to save patients from unneeded exposure to medical imaging with a smaller number of m views of even smaller number k of radiation exposing pixels as ones passing a filter among zeros. Thus, CS is not post-processing image compression, because otherwise, the patients have already suffered, and CS happened before the sensing measurement. They adopted a purely random sparse sampling mask $[\Phi]_{m,N}$ consisting of $m \cong k$ number of one's (passing the radiation) among seas of zeros (blocking the radiation). The goal of multiplying such a long horizontal rectangular sampling matrix $[\Phi]_{m,N}$ is to achieve the linear dimensionality reduction from N to m (m<<N), and the reduced square matrix follows:

$$\vec{y}_m = [B]_{m,m}\vec{X}_m, \qquad (2)$$

where $\vec{y}_m \equiv [\Phi]_{m,N}\vec{Y}_N$ and $[B]_{m,m} \equiv [\Phi]_{m,N}[A]_{N,m}$. Remarkably, given a set of sparse orthogonal measurements $\vec{y}_m$, they reproduced the original resolution medical image $\vec{X}_N$. CRT&D used an *iterative hard*

threshold (IHT) of the largest guesstimated entries, known as linear programming, based on the $min.|\vec{X}|_1$ subject to $E = |\vec{y}_m - [B]_{m,m}\vec{X}_m|_2^2 \le \varepsilon$, where l_p-norm is defined $\left|\vec{X}\right|_p \equiv (\sum_{n=1}^{N}|x_n|^p)^{1/p}$.

ANN supervised learning adopts the same LMS errors. Energy between the desired outputs $\vec{v}'_i = \vec{y}_m$ & the actual weighted output $v_i = \sigma(u_i) = \sigma(\sum_{j=1}^{m}[W_{i,j}]u_j)$. Neurodynamics I/O is given $du_i/dt = -\partial E/\partial v_i$; Lyapunov convergence theorem $dE/dt \le 0$ is proved for the monotonic sigmoid logic $d\sigma/du_i \ge 0$ in Sect.2. ANN does not use the Manhattan distance, or going around a city-block l_1 distance at $p = 1$ because it is known in ANN learning to be too sensitive to outliers. Mathematically speaking, the true sparseness measure is not the l_1-norm but the l_0-norm: $\left|\vec{X}\right|_0 \equiv (\sum_{n=1}^{N}|x_n|^0)^{1/0} = k$, counting the number of non-zero elements because only non-zero entry raised to zero power is equal to 1. Nevertheless, a practical lesson from CRT&D is that the $min.|\vec{X}|_1$ subject to $min.|\vec{X} - \vec{Y}|_2^2$ is sufficient to avoid the computationally intractable $min.|\vec{X}|_0$. In fact, without the constraint of minimization l_1 norm, LMS is blind to all possible direction cosines within the hyper-sphere giving rise to the *Penrose's inverse* $[A]^{-1} \cong [A]^T([A][A]^T)^{-1}$; or $[A]^{-1} \cong ([A]^T[A])^{-1}[A]^T$ (simplified by A=QR decomposition). To be sure, CRT&D proved a Restricted Isometry Property (*RIP*) Theorem, stating that a limited bound on a purely random sparse sampling: $\| [\Phi]_{m,N}\vec{X} \|/\| \vec{X} \| \cong O(1 \mp \delta_k)$, $m \cong 1.3k \ll N$. As a result, the $min.l_1$-norm is equivalent to the $min.l_0$-norm at the same random sparseness. Such an equivalent linear programming algorithm takes a manageable polynomial time. However, it is not fast for video imaging.

The **subtitles** may help those who wish to selectively read about ANN and NI. The universal language is mathematics.

NI Definition: NI may be defined by unsupervised learning algorithms running iteratively on connectionist architectures, naturally satisfying fault tolerance (FT), and subspace generation (SG).

Hebb Learning Rule: If blinking traffic lights at all street intersections have built-in data storage from all the transceivers, then traffic lights function like neurons with synaptic junctions. They send and receive a frequency modulation Morse code ranged from 30 Hz to 100 Hz firing rates. Physiologist Donald Hebb observed the plasticity of synaptic junction learning. The Hebbian rule describes how to modify the traffic light blinking rate to indicate the degree of traffic jam at street intersections. The plasticity of synapse matrix $[W_{i,j}]$ is adjusted in proportion to the inputs of u_i from the i-th street weighted by the output change, Δv_j at the *j*-th street as the vector outer product code:

$$\text{Do 10:} \quad \Delta W_{j,i} \approx \Delta v_j\, u_i \ . \tag{3a}$$

$$10: \quad [W_{j,i}]' = [W_{j,i}] + \Delta W_{j,i}; \text{ Return.} \tag{3b}$$

An event is represented in the m-subspace of *N*=10 billion neurons in our brain by the synergic blinking pattern of *m* communicating neurons/traffic lights. The volume of m-subspace may be estimated by the vector outer products called associative memory [AM] matrix inside the hippocampus of our central brain (Fig. 4c). Even if a local neuron or traffic light broke down, the distributed associative memory (AM) can be retrieved. This is the FT as the nearest neighbor classifier in a finite solid angle cone around each orthogonal axis of the subspace; then, the subspace generalization is going along a new direction orthogonal to the full subspace.

Unsupervised lesson learned: Supervised learning stops when the algorithm has achieved a desired output. Without knowing the desired output, an unsupervised learning algorithm doesn't know when to stop. Since the input data already has some energy in its representation; the measurement principle should not bias the input energy for firing sensory system reports accordingly. Thus, the magnitude of output's firing rates should not be changed by the learning weight. Note that in physics the photon energy field is the quadratic displacement of oscillators. Thus, the constraint of unsupervised learning requires adjusting the unit weight vector on the surface of hyper-sphere of R^m. In fact, the main lesson of Bell-Sejnowski-Amari-Oja (BSAO) unsupervised learning algorithm is this natural stopping criterion for the given set of input vectors $\{\vec{x}\} \in R^N$. The BSAO projection pursuit algorithm is merely a rotation within a Hyper-sphere. It stops when the weight vectors $[W_{i,j}] = [\vec{w}_i, \vec{w}_i, ..]$ becomes parallel in time to the input vectors of any magnitude $\vec{w}_i||\vec{x}_i$. The following theorem will be derived thrice in Sect.2

$$\left[\boldsymbol{\delta_{\alpha,\beta} - x_\alpha x^T{}_\beta}\right]\vec{x}_\alpha\vec{x}_\beta = \mathbf{0},$$

$$\Delta\vec{w} \equiv \vec{w}' - \vec{w} = \boldsymbol{\epsilon}\left[\boldsymbol{\delta_{\alpha,\beta} - \vec{w}'_\alpha \vec{w}'^T{}_\beta}\right]\vec{x}_\alpha \frac{dK(\vec{u}_\beta)}{d\vec{w}_i}\ , \tag{3c}$$

where K is a reasonable source separation contrast function of the weighted input $\vec{u}_i = [W_{i,\gamma}]\vec{x}_\gamma$. The contrast function could be (i) the maximum a-posteriori entropy (filtered output entropy) used in Bell & Senowski algorithm in 1996; (ii) the fixed point algorithm of 4-th order cumulant Kurtosis (Fast ICA) adopted in Hyvarinen & Oja in 1997; (iii) the isothermal equilibrium of minimum thermodynamic Helmholtz free energy ($Brain\ T_o = 37^o\,C$) known as Lagrange Constraint Neural Net, in terms of $min.H = E - T_o\, max.S$ (maximum a-priori source entropy) by Szu & Hsu, 1997.

When we were young, unsupervised learning guided us to effortlessly extract sparse orthogonal neuronal representations in a minimum isothermal free energy fashion. Subsequently, the supervised learning expert system at school comes in handy with these mental representations. As we get older, our unsupervised ability for creative *e*-Brain is eroded and outweighed by the expert system *l*-Brain.

In this paper, we assume without rigorous proof that the RIP theorem works for both the purely random sparse

$[\Phi]_{m,N}$ and the organized sparse $[\Phi_s]_{m,N}$. A sketch of proof using exchange entropy for the complexity of organized sparseness is given at the end of Sect. 7. Intuitively speaking, we do not change the quantity of ones of $[\Phi_s]_{m,N}$; only we endow a feature meaning to the ones' locations beyond CS all pass filtering. In other words, the admissible ANN storage demands the orthogonal sparse ***moments of*** spotting ***dramatic orthogonal changes of salient features.*** Consequently, we will not alter the value of unknown image vector $\| \vec{X} \|$ more than δ_k. In fact, for real-time video, we have bypassed random CS coding and image recovery algorithms, and chose instantaneous retrieval by MPD Hetro-Associative Memory (HAM) storage defined in Sections 2 and 4.

After the introduction of the goals and the approaches of CS, we review ANN in Section 2; Neuroscience 101 with an emphasis on the orthogonal sparseness representations of HVS in Section 3; the AM storage in Section 4; Software simulation results in Sect. 5; Unsupervised Spatial-Spectral CS Theory in Section 6; and Hardware design of camera in Section 7. This review might provide the first mathematical theory of learning from supervised ANN to unsupervised ANN.

2. Reviews of Artificial Neural Networks

Traditional ANN performs supervised learning, known as a soft lookup table, or merely as 'monkey sees, monkey do'. Marvin Minskey and Seymour Papert commented on ANN and Artificial Intelligence (AI) as well as extending multi-layer finite state machine to do "Ex OR" in 1988. This was about the year when INNS was incepted by 17 interdisciplinary Governors. Notably, Stephen Grossberg, Teuvo Kohonen, Shun-ichi Amari serve as editors in chief of INNS, which was published quarterly by Pergamum Press and subsequently, monthly by Elsevier Publisher. In the last two decades, both INNS and Computational Intelligence Society of IEEE have accomplished a lot (recorded in IJCNN proceedings). Being the founding Secretary and Treasurer, a former President and Governor of INNS, the author apologizes for any unintentional bias. Some opinions belong to all shade of grey; taking binary or spicy story approach is one of the great pedagogical techniques that our teachers George Uhlenbeck and Mark Kac often did at the Rockefeller University. For example, 'nothing wrong with the supervised learning exemplars approach using the lookup table, the curse is only at the 'static' or closed lookup table.' Also, 'this limitation of supervised learning is not due to the connectionist concept, rather, due to the deeply entrenched "near equilibrium" concept'; Norbert Wiener developed near equilibrium Cybernetics in *1948 & 1961*. 'What's missing is the ability to create a new class far away from equilibrium.' 'INNS took the out of the box, interdisciplinary approach to learn from the Neurosciences how to develop unsupervised learning paradigm from Neurobiology.' 'This is an important leg of NI tripod. The other two legs are the fault tolerance and the subspace generalization.'

2.1 Fault Tolerance and Subspace Generalization

Fault Tolerance (FT): The read out of m neuronal representation satisfies the fault tolerance. This is due to the geometry of a circular cone spanned in 45^o solid angle around the m-D vector axis. This central axis is defined as the memory state and the cone around it is its family of turf vectors. Rather than precisely pointing in the same vector direction of m-D, anything within the turf family is recognized as the original axis. This is the reason that the read out is fault tolerant. Thus, [AM] matrix storage can enjoy a soft failure in a graceful degradation fashion, if and only if (iff) all storage state vectors are mutually orthogonal within the subspace; and going completely outside the subspace in a new orthogonal direction to all is the subspace generalization (SG).

Subspace Generalization (SG): We introduce the inner product bracket notation $< Bra|Ket >= c$, in the dual spaces of $< Bra|$ and $|ket >$, while the outer product matrix is conveniently in the reverse order $[w_{j,i}] = |v_j >< u_i|$ introduced by physicist P. Dirac. We prove the 'traceless outer product' matrix storage allows SG from the m-D subspace to one bigger m+1-D subspace. Defined, the Ortho-Normal (ON) basis is $< n'|n >= \delta_{n',n}$; $n, n' = 1,..m$. Then, SG is the Trace-less ON $[AM]_{m,m} = \sum_{n=1}^{m} |n >< n| - Tr[AM]_{m,m}$. Trace operator Tr: summing all diagonal elements is the projection operator defined $Tr^2 = Tr$.

SG Theorem: Without supervision, a traceless matrix storage of ON sub-space can self-determine admitting $|x > = |m+1 >$, iff $< m+1|n >= \delta_{m+1,n}$ satisfying the fixed point of cycle 2 rule: $[AM]_{m,m}{}^2|x > = |x >$, then $[AM]_{m+1,m+1} = \sum_{n=1}^{m+1} |n >< n| - Tr[AM]_{m+1,m+1}$.

Q.E.D.

AM is MPD computing, more than the nearest neighbor Fisher classifier. These FT & SG are trademarks of connectionist, which will be our basis of CS approach. Unsupervised learning is a dynamic trademark of NI. New learning capability comes from two concepts, (i) engineering filtering concept and (ii) physics-physiology isothermal equilibrium concept.

Semantic Generalization: Semantic generalization is slightly different than the subspace generalization, because it involves a higher level of cognition derived from both sides of the brain. Such an e-Brain and *l*-Brain combination processes thinking within two boxes of brain related by a set of independent degrees of freedom. Thus, this semantic generalization is the different side of the same coin, in Sect 4. We are ready to set up the math language leading to the modern unsupervised learning as follows:

2.2 Wiener Auto Regression

Norbert Wiener invented the *near equilibrium control* as follows. He demonstrated a negative feedback loop for the missile trajectory guidance. He introduced a moving average Auto Regression (AR) with LMS error:

$$min.E =< (u_{(m)} - y)^2 >$$

where the scalar input $u_{(m)} = \vec{w}_m{}^T \vec{x}_m \equiv < \vec{x}_m >$ has weighted average of the past *m* data vector

$$\vec{x}_m = (x_t, x_{t-1}, x_{t-2}, \dots, x_{t-(m-1)})^T$$

to predict the future as a desired output $y = x_{m+1}$.

A simple near equilibrium analytical filter solution is derived at the fixed point dynamics

$$\frac{\partial E}{\partial \vec{w}} = 2 < (\vec{w}^T \vec{x}_m - x_{m+1})\vec{x}_m >= 0, \quad \text{e.g.} \quad m=3$$

$$\begin{bmatrix} c_o & c_1 & c_2 \\ c_1 & c_o & c_1 \\ c_2 & c_1 & c_o \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix};$$

$c_{t-t'} \equiv < x_t x_{t'} >;\ c_1 =< x_t x_{t-1} >; c_2 =< x_t x_{t-2} >; \dots$

Solving the Teoplitz matrix, Wiener derived the filter weights. Auto Regression (AR) was extended by Kalman for a vector time series with nonlinear Riccati equations for extended Kalman filtering. In image processing: $\vec{X} = [A]\vec{S} + \vec{N}$ where additive noisy images become a vector $\vec{X}$ represented by a lexicographic row-by-row order over 2-D space $\vec{x}$. Wiener image filter is derived using AR fixed point (f.p.) algorithm in the Fourier transform domain:

$$\hat{X}(\vec{k}) = \iint d^2\vec{x} \exp(j\vec{k} \cdot \vec{x})\, \vec{X}(\vec{x}); j = \sqrt{-1}$$

Using Fourier de-convolution theorem, $\exp(j\vec{k} \cdot \vec{x}) = exp(j\vec{k} \cdot (\vec{x} - \vec{y}))\exp(j\vec{k} \cdot \vec{y}])$, gives a linear image equation in the product form in Fourier domain:

$$\hat{X} = \hat{A}\hat{S} + \hat{N}$$

Wiener sought $\hat{S} = \widehat{W}\hat{X}$ to minimize the LMS errors

$$E =< \left(\widehat{W}\hat{X} - \hat{S}'\right)^* \cdot \left(\widehat{W}\hat{X} - \hat{S}'\right) >;$$
$$\because f.p.\ \frac{\partial E}{\partial \widehat{w}^*} =< 2\hat{X}^* \cdot \left(\widehat{W}\hat{X} - \hat{S}'\right) >= 0;$$

Termination Condition: if $\overrightarrow{data} \perp \overrightarrow{error} \to 0: \hat{S}' \to \hat{S}$

$$\therefore \widehat{W} =< \hat{X}^* \cdot \hat{S}' > [< \hat{X}^* \cdot \hat{X} >]^{-1} \cong \hat{A}^{-1}[1 + \varepsilon]^{-1},$$

where noise to signal ratio $\varepsilon \equiv < \hat{N}^* \hat{N} > / |\hat{A}|^2 |\hat{S}|^2$.

Wiener filtering is the inverse filtering $\widehat{W} = \hat{A}^{-1}$ at strong signals, and becomes Vander Lugt filtering $\widehat{W} = \hat{A}^* \frac{|\hat{S}|^2}{<|\hat{N}|^2>}$ for weak signals. A mini-max filtering is given by Szu (Appl. Opt. V. 24, pp.1426-1431, 1985).

Such a near equilibrium adjustment influenced generations of scientists. While F. Rosenblatt of Cornell U. pioneered the 'perceptron' concept for OCR, B. Widrow of Stanford took a leap of faith forward with 'multiple layers perceptrons.' Hyvarinen & Oja developed the Fast ICA algorithm. The author was fortunate to learn from Widrow; co-taught with him a short UCLA course on ANN, and continued teaching for a decade after 1988 (thanks to W. Goodin).

2.3 ANN generalize AR

Pedagogically speaking, ANN generalizes Wiener's AR approach with 4 none-principles: (i) non-linear threshold, (ii) non-local memory, (iii) non-stationary dynamics and (iv) non-supervision learning, respectively Equations (4a,b,c,d).

2.3.1 Non-linear Threshold: Neuron model McCullouch & Pitts proposed in 1959 a sigmoid model of threshold logic: mapping of neuronal input $u_i(-\infty, \infty)$ to the unary output $v_i[0,1]$ asymptotically by solving Ricati nonlinear $\frac{dv_i}{du_i} = v_i(1 - v_i) = 0$, at 'no or yes' limits $v_i = 0;\ v_i = 1$. Exact solution is:

$$v_i = \sigma(u_i) \equiv [1 + \exp(-u_i)]^{-1}$$
$$= \exp\left(\frac{u_i}{2}\right)[\exp\left(\frac{u_i}{2}\right) + \exp\left(-\frac{u_i}{2}\right)]^{-1}, \qquad (4).$$

Three interdisciplinary interpretations are given:

Physics, this is a two state equilibrium solution expressed in firing or not, the canonical ensemble of the brain at the equilibrium temperature T, and the Boltzmann's constant K_B, as well as an arbitrary threshold θ: $y = \sigma Tx - \theta = 1 + \exp -x - \theta KBT - 1.$

Neurophysiology, this model can contribute to the binary limit of a low temperature and high threshold value a single ***grandmother*** neuron firing in a family tree subspace (1,0,0,0,0..) as a ***sparse network representation.***

Computer Science, an overall cooling limit, $K_B T \Longrightarrow 0$, the sigmoid logic is reduced to the binary logic used by John von Neumann for the digital computer: $1 \geqq \sigma_o(x \geq \theta) \geqq 0$.

2.3.2 Nonlocal memory: D. Hebb learning rule of the communication is efficiently proportional to what goes in and what comes out the channel by $W_{i,j} \propto v_i\, u_j$ measuring the weight matrix of inter-neuron synaptic gap junction. A weight summation of $\vec{x}_i$ given by **Compressive Sensing** rise to a potential sparse input $\vec{u}_i = [W_{i,\alpha}]\vec{x}_\alpha$ Eq(3).

2.3.3 Non-stationary dynamics: Laponov control theory insures the convergence of neurodynamics $\frac{du_i}{dt} = -\frac{\partial E}{\partial v_i}$.

2.3.4 Non-supervised Learning Nonconvex energy landscape: $E \cong H(open\ set\ of\ exemplars)$.

2.4 Energy Landscape Supervised Algorithm

Physicist John Hopfield broadened the near-equilibrium engineering notion and introduced the mental energy of a non-convex landscape E(v_i) at the output v_i space to accommodate the (neurophysiologic known) associative memory storage. He introduced Newtonian dynamics $du_i/dt = -\partial E/\partial v_i$ as a generalization of the fixed point LMS solution. He proved a control theory Lyapunov convergence:

$$\frac{dE}{dt} = \sum_i \frac{\partial E}{\partial v_i}\frac{dv_i}{du_i}\frac{du_i}{dt} = -\sigma'\left(\frac{\partial E}{\partial v_i}\right)^2,$$

independent of energy landscapes, as long as a real monotonic positive logic $dv_i/du_i \equiv \sigma' \geq o$, in terms of $(\text{in}, \text{out}) = (u_i, v_i)$ defined by

$$v_i = \sigma(u_i); \& \, u_i = \sum_j W_{i,j} v_j \, ; E = -\tfrac{1}{2} \sum_{i,j} W_{i,j} v_i v_j .$$

Physicist E. R. Caianiello is considered a thinking machine beyond Wiener's AR. He used causality physics principles to generalize the instantaneous McCullough & Pitts neuron model building in the replenishing time delay in 1961.

Psychologist James Anderson, in 1968, developed a correlation memory while **Christopher von der Malsburg,** 1976, developed a self-organization concept. They described a *brain in a box* concept, inspired by the binary number predictor box built by **K. Steinbuch & E. Schmidt** and based on a *learning matrix* as the beginning of *Associative Memory* (AM) storage in biocybernetics in Avionics 1967. **Kaoru Nakano,** 1972, and **Karl Pribram,** 1974, enhanced this distributive AM concept with a fault tolerance (FT) for a partial pattern completion (inspired by Gabor hologram).

Engineer Bernard Widrow took multiple layers perceptrons as adaptive learning neural networks. For computing reasons, the middle layer neurons took the cool limit $T \to 0$ of the sigmoid threshold as non-differentiable bipolar logic, and achieved a limited adaptation.
From the connectionist viewpoints, **Shun-ichi Amari** indicated in 1980 that the binary logic approach might suffer a premature locking in the corners of hyper-cubes topology.

2.5. Backprop Algorithm

It took a team of scientists known as the Cambridge PDP group (Neuropsychologists David Rumelhart, James McClelland, Geoffrey Hinton, R. J. Williams, Michael Jordan, Terrence Sejnowski, Francis Crick, and graduate students) to determine the backprop algorithm. They improved Wiener's LMS error $E = \sum_i |v_i - v_i^*|^2$ with a parallel distributed processing (PDP) double decker hamburger architecture, consisting of 2 layers of beef (uplink $w_{k,j}$ & downlink $w'_{j,i}$) sandwiched between in 3 layers of buns made of neurons. The sigmoid logic: $\sigma' \equiv d\sigma/du_k < \infty$ is analytic, they unlocked the bipolar 'bang-bang' control from Widrow's corners of hypercubes. They have analytically derived the 'Backprop' algorithm. Namely, passing boss error to that of the hidden layer; and, in turns, to the bottom layer which has exemplar inputs.

$$\frac{\partial w_{j,i}}{\partial t} \cong \frac{\Delta w_{j,i}}{\Delta t} = -\frac{\partial E}{\partial w_{j,i}} \tag{5a}$$

The Hebb learning rule of uplink weight is obtained by the chain rule:

$$\Delta w_{k,j} = -\frac{\partial E}{\partial w_{k,j}} \Delta t \cong -\sum_n \frac{\partial E}{\partial u_n} \frac{\partial u_n}{\partial u_k} \frac{\partial u_k}{\partial w_{k,j}} \Delta t$$
$$= \sum_n \delta_n \delta_{n,k} v'_j \Delta t = \delta_k v'_j \Delta t, \tag{5b}$$

Kronecker $\delta_{n,k} \equiv \frac{\partial u_n}{\partial u_k}$ selects top layer post-synaptic δ_j (error energy slope) and hidden layer pre-synaptic v'_i:

$$\delta_k \equiv -\frac{\partial E}{\partial u_k} = -\frac{\partial E}{\partial v_k} \frac{\partial v_k}{\partial u_k} = -(v_k - v_k^*)\sigma^{(\prime)}. \tag{5c}$$

The sigmoid slope $\sigma^{(\prime)}$ is another Gaussian-like window function. The PDP group assumed Hebb's rule $\Delta w_{k,j} \approx \delta_k v'_j$ holds true universally, and cleverly computed the hidden share of blaming δ'_j from fan-in boss errors δ_k

$$\delta'_j \equiv -\frac{\partial E}{\partial u'_j} = -\sum_k \frac{\partial E}{\partial u_k} \frac{\partial u_k}{\partial v'_j} \frac{\partial v'_j}{\partial u'_j} \equiv \sum_k \delta_k w_{k,j} \sigma^{(\prime)}. \tag{5d}$$

Each layer's I/O firing rates are denoted in the alphabetic order as (input, output) = (u,v) respectively; the top, hidden, and bottom layers are labeled accordingly:

$$(v_k, u_k) \leftarrow w_{k,j} \leftarrow (v'_j, u'_j) \leftarrow w'_{j,i} \leftarrow (v"_i, u"_i),$$

where $v_k = \sigma(u_k) \equiv \sigma(\sum_j w_{k,j} v'_j)$; $v'_j = \sigma(u'_j \equiv \sum_i w'_{j,i} v''_i)$. Hebbian rule turns out to be similar at every layers, e.g., $\delta''_j \equiv -\frac{\partial E}{\partial u''_j} = \sum_k \delta'_k w'_{k,j} \sigma^{(\prime)}$, etc. Such a self-similar chain relationship is known as backprop.

Bio-control: Independently, Paul Werbos took a different viewpoint, assigning both the delta of credit and the delta of blame to the performance metric at different locations of the feedback loops in real world financial-like applications. As if this were a carrot and stick model controlling a donkey, to be effective, these feedback controls must be applied at different parts of the donkey. Thus, this bio-control goes beyond the near-equilibrium negative feedback control. Such thinking began a flourishing era, notably, Kumpati Narendra, et al. produced stochastic, chaos, multi-plants, multi-scales, etc., control theories.

2.6 Self-Organization Map (SOM)

Teuvo Kohonen computed the batched centroid update rule sequentially:

$$< \vec{x} >_{N+1} = < \vec{x} >_N \left(\frac{N+1-1}{N+1}\right) + \frac{1}{N+1} \vec{x}_{N+1}$$
$$= < \vec{x} >_N + \rho(\vec{x}_{N+1} - < \vec{x} >_N), \tag{6}$$

replacing the uniform update weight with adaptive learning $\rho = \frac{1}{N+1} < 1$. SOM has contributed to database applications with annual world-wide meetings, e.g. US PTO Patent search, discovery of hidden linkage among companies, genome coding, etc.

2.7 NP Complete

David Tank and John Hopfield (T-H) solved a class of computationally intractable problems (classified as the nondeterministic polynomial (NP), e.g. the Travelling Salesman Problem, Job scheduling, etc.) The possible tours are combinatorial explosive in the factorial $N!/2N$, where the denominator is due to the TSP having no home base, and the clockwise and counter-clockwise tours having an equal distance. T-H solved this by using the powerful MPD

computing capability of ANN (Cybernetics, & Sci. Am. Mag).

$$E = \sum_{\vec{c}=1}^{N} \sum_{\vec{t}=1}^{N} v_{\vec{c}} \left[W_{\vec{c},\vec{t}}\right] v_{\vec{t}} + \text{Constraints.}$$

Their contribution is similar to DNA computing for cryptology RSA de-coding. Both deserve the honor of Turing Prizes. Unfortunately, the T-H constraints of the permutation matrix $\left[W_{\vec{c},\vec{t}}\right]$ are not readily translatable to all the other classes of the NP complete problems:

$$\left[W_{\vec{c},\vec{t}}\right] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

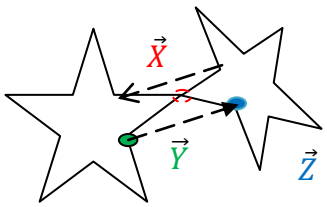
(city #1 visits tour No.1; #2 for 2nd , #3 for 3rd etc., and returned to No. 1; T-H labeled each neuron with 2-D vector index for convenience in both city & tour indices. Meanwhile, Y. Takefuji & others mapped the TSP to many other applications including genome sequencing (Science Mag.).

Divide & Conquer (D&C) Theorem: Szu & Foo solved a quadratic NP complete computing by D&C, using orthogonal decompositions $\vec{A} = \vec{B} + \vec{C}$, l_2-norm:

$$min.|\vec{A}|_2^2 = min.|\vec{B}|_2^2 + min.|\vec{C}|_2^2;\ iff\ \vec{B} \cdot \vec{C} = 0. \qquad (7)$$

Unfortunately, searching boundary tours for divisions could be time consuming. Moreover, Simon Foo and I could not solve the TSP based on the original constraints of row-sum and column-sum and row products and column products of the matrix, generating a Mexican standoff dual like in Hollywood western movies.

Improved TSP with D&C Algorithm: A necessary and sufficient constraint turns out to be *sparse orthogonal sampling Matrix* $[\Phi_s]$ which is equivalent to a permutation mix up of the identity matrix. Iff each row and column is added up to one, similar to N queens constraint (in chess, queens can kill each other, unless only one queen occupies one row and column). Furthermore, for a Go game for gaining territory one puts down a white stone in the center of the board about a half size creating a surrogate or ghost city at the point $\vec{X}$.



Without the need of a boundary search among cities, adding a ghost city $\vec{X}$ finds two neighborhood cities $\vec{Y}$ and $\vec{Z}$ with two vector distances: $\vec{B} = \vec{Y} - \vec{X}$, $\vec{C} = \vec{X} - \vec{Z}$. Iff $\vec{B} * \vec{C} = 0$ satisfies the D&C theorem, we accept $\vec{X}$. Then we conceptually solve two separate TSP problems in parallel. Afterwards, we can remove the ghost city and modify the tour sequences indicated by dotted lines. According to the triangle inequality, $|\vec{a}| + |\vec{b}| \geq |\vec{c}|$, the vector $\vec{c}$ represents a dotted line having a shorter tour path distance than the original tour involving the ghost city. **Q.E.D.**

This strategy should be executed from the smallest doable regions to bigger ones; each time one can reduce the computational complexity by half. In other words, solving the total N=18 cities by two halves N/2=9; one continues the procedure without stopping solving them, further dividing 9 by 4 and 5 halves, until one can carry out TSP in smaller units 4 and 5, each has de-ghosted afterward. Then, we go on to 9 and 9 cities, and de-ghost in a reverse order.

2.8 ART

Gail Carpenter and Stephen Grossberg implemented the biological vigilance concept in terms of two layers of analog neurons architecture. They proved a convergence theorem about short and long term traces $Z_{i,j}$ in 1987 App. Op. Their two layer architecture could be thought of as if the third layer structure were flipping down to touch the bottom layer using two phone lines to speak to one another top-down or bottom up differently. Besides the PDP 3 layers buns sandwiched 2 layers of weights have the original number of degrees of freedom, they created a new degree of freedom called the vigilance defined by $\rho = (\vec{m}_{t+1}, < \vec{w}_t >) = cos(\leq \pi/4) \geq 0.7$. This parameter can either accept the newcomer and updating the leader's class Centroid with the newcomer vector; or rejecting the newcomer letting it be a new leader creating a new class. Without the need of supervision, they implemented a self-organizing and robust MPD computing who follows the leader called (respectively binary, or analog, or fuzzy) the adaptive resonance theory (ART I, II, III). ART yields many applications by Boston NSF Center of Learning Excellence. M. Cader, et. al. at the world bank implemented ART expert systems for typing seeds choice for saving the costly diagnosis needs by means of PCR amplification in order to build up enough DNA samples (pico grams); a decision prediction system based on the past Federal Reserve open forum reports (Neural Network Financial Expert Systems, G. J. Deboeck and M. Cader, Wiley 1994).

2.9 Fuzzy Membership Function

Lotfi Zadeh introduced an open set for imprecise linguistic concept represented by a 'possibilities membership function', e.g. beauty, young, etc. This open set triangular shape function is not the probability measure which must be normalized to the unity. Nevertheless, an intersection of two fuzzy sets, e.g. young and beautiful, becomes sharper in the joint concept. Such an electronic computing system for the union and the intersection of these triangle membership functions is useful, and has been implemented by Takeshi Yamakawa in the fuzzy logic chips. Whenever a precise engineering tool meets an imprecise application, the fuzzy logic chip may be useful, e.g. automobile transmission box and household comfort control device. Such a fuzzy logic technology becomes exceedingly useful as documented in a decade of soft computing conferences sponsored by The Japan Fuzzy Logic Industry Association.

ANN Modeling of Fuzzy Memberships: Monotonic sigmoid logic is crucial for John Hopfield's convergence proof. If the neuron had a piecewise negative response in the shape of a scripted N-letter: $v_i = \sigma_N(u_i)$, then, like the logistic map, it has a single hump height adjustable by the λ-knot (by M. Feigenbaum for the tuning of period doubling bifurcation cascade). If we represent each pixels by the sick neuron model $v_i = \sigma_N(u_i)$, then recursively we produce the nonlinear Baker transform of image mixing. Such a Chaotic NN is useful for the modeling of drug-induced hallucinating images, olfactory ball smell dynamics of Walter Freeman, and learnable fuzzy membership functions of Lotfi Zadeh.

2.10 Fast Simulated Annealing

Szu and Hartley have published in Phys Lett. and IEEE Proc. 1986, the Fast Simulated Annealing (FSA) approach. It combines the increasing numbers of local Gaussian random walks at a high temperature *T*, with an unbounded Levy flights at a low temperature in the combined Cauchy probability density of noise. A speed up cooling schedule is proved to be inversely linear time step $T_C = \frac{T_o}{1+t'}$, for any initial temperature T_o that guarantees the reaching of the equilibrium ground state at the minimum energy. Given a sufficient low temperature $\tilde{T}_o$ Geman and Geman proved in 1984 the converging to the minimum energy ground state by an inversely logarithmic time step: $T_G = \frac{\tilde{T}_o}{1+log(1+t)}$. Sejnowski & Hinton used the Gaussian random walks in the Boltzmann's machine for a simulated annealing learning algorithm emulating a baby learning the talk, called Net-talk.

Cauchy Machine: Y. Takefuj & Szu designed an electronic implementation of such a set of stochastic Langevin equations. Stochastic neurons are coupled through the synapse AM learning rule and recursively driven by Levy flights and Brown motions governed by the Cauchy pdf. The set of Cauchy-Langevin dynamics enjoys the faster inversely linear cooling schedule to reach the equilibrium state.

Do 10 t'=t'+1; $T_C(t') = \frac{T(t\prime)}{1+t'}$
$\Delta x = T_c(t')\ \tan[(2\theta[0,1] - 1)\ \pi/2]$;
x(t')=$x_{t\prime}$+Δx; $E(t') = \sum_{i=1}^{m} \frac{1}{2} k(x(t') - x_i)^2$;
$\Delta E = E(t') - E(t'-1)$;
If $\Delta E \leq 0$; accept x(t'), Go To 10;
or, compute $\exp(-\Delta E/K_B T_C(t')) > \varepsilon_o$;
accept x(t')
10: Return.

Optical version of a Cauchy machine is done by Kim Scheff and Joseph Landa. The Cauchy noise is optically generated by the random reflection of the mirror displacement *x* of the optical ray from a uniformly random spinning mirror angle $\theta(-\frac{\pi}{2}, \frac{\pi}{2})$. The temperature T is the distance parameter between the mirror and the plate generates the Cauchy probability density function (pdf) (Kac said as a French counter example to the British Gauss Central Limiting Theorem). This pdf is much faster for search than Gaussian random walks:

$$\rho_G(\Delta x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\Delta x^2}{T}\right) \cong \frac{1}{\sqrt{2\pi}}(1 - \frac{\Delta x^2}{T} + ..).$$
$$\rho_C(\Delta x) = \frac{1}{\pi}(1 + \frac{\Delta x^2}{T})^{-1} = \frac{1}{\pi}(1 - \frac{\Delta x^2}{T} + \cdots).$$

Proof: $\frac{dx}{d\theta} = T(1 + tan\theta^2); \pi = \int d\theta = \int \frac{d(\frac{x}{T})}{1+tan\theta^2}$;
$1 = \int \rho_c(x)dx = \frac{1}{\pi}\int \frac{1}{1+(\frac{x}{T})^2} d(\frac{x}{T})$. *Q.E.D.*

Global Levy Flights $< \Delta x^2 >_{\rho_C} = \infty$
Local Brownian motion $< \Delta x^2 >_{\rho_C} \cong T(t)$

NI Expert System: Szu and John Caulfield published an optical expert system in 1987, generalized the AI Lisp programming the pointer linkage map from 1-D vector arrays of $\vec{f} = (A,\ O, V)^T$ to $\vec{f}' = (A',\ O',\ V')^T$, etc. The color, "A attribute," of apple, "O object," is red, "V value". We represent the Lisp map with the MPD $[HAM] = \sum \vec{f}\,\vec{f}'^T$ storage which has demonstrated both the FT and the Generalization capabilities.

2.11 Unsupervised Learning of *I*-Brain

In order to make sure nothing but the desired independent sources coming out of the filter, C. Jutten and J. Herault adjusts the weights of inverse filtering to undo the unknown mixing by combining the inverse and forward operation as the unity operator (Snowbird IOP conf.; Sig. Proc. 1991). Since J.-F. Cardoso has systematically investigated the blind de-convolution of unknown impulse response function; he called a matrix form as Blinded Sources Separation (BSS) by non-Gaussian higher order statistics (HOS), or the information maximum output. His work since 1989 did not generate the excitement as it should be in the ANN community. It was not until Antony J. Bell and Terry J. Sejnowski (BS), et al. [10] have systematically formulated an *unsupervised learning of ANN algorithm*, solving both the unknown mixing weight matrix and the unknown sources. Their solutions are subject to the constraints of maximum filtered entropy $H(y_i)$ of the output $y_i = [W_{i,\alpha}]x_\alpha$, where $x_j = [A_{j,\alpha}?]s_\alpha?$, and the repeated Greek indices represent the summation. ANN model uses a robust saturation of linear filtering in terms of a nonlinear sigmoid output $y_i = \sigma(x_i) = \{1 + \exp(-[W_{i,\alpha}]x_\alpha)\}^{-1}$. Since a single neuron learning rule turns out to be isomorphic to that of *N* neurons in tensor notions, for simplicity sake we derive a single neuron learning rule to point out why the engineering filter does not follow the Hebb's synaptic weight updates. Again, a bona fide unsupervised learning does not need to prescribe the desirable outputs for exemplars inputs. For ICA, BS chose to maximize the Shannon output entropy H(y) indicating that the inverse filtering has blindly de-convoluted and found the unknown independent sources without knowing the impulse response function or mixing matrix. Thus, the filter weight adjustment is defined in the following and the BS result is derived as follows:

$$\frac{\delta w}{\delta t} = \frac{\partial H(y)}{\partial w}, \&,\ H(y) = -\int f(y)\log f(y)dy \Longrightarrow$$

$$\delta w = \frac{\partial H(y)}{\partial w}\delta t = \{|w|^{-1} + (1-2y)x\}\delta t. \qquad (8a)$$

Derivation: From the normalized probability definitions:

$$\int f(y)dy = \int g(x)dx = 1\,; f(y) = \frac{g(x)}{\left|\frac{dy}{dx}\right|};$$

$$H(y) \equiv -<\log f(y)>_f\ ,$$

we express the output pdf in terms of the input pdf with changing Jacobian variables. We exchange the orders of operation of the ensemble average brackets and the derivatives to compute

$$\frac{\partial H(y)}{\partial w} = \frac{\partial <\log\left|\frac{dy}{dx}\right|>_f}{\partial w} \cong |\frac{dy}{dx}|^{-1}\frac{\partial\left|\frac{dy}{dx}\right|}{\partial w};$$

Ricati sigmoid: $y = [1+\exp(-wx)]^{-1}$;

$$\frac{dy}{d(wx)} = y(1-y); \frac{dy}{dx} = wy(1-y) \& \frac{dy}{dw} = xy(1-y).$$

Substituting these differential results into the unsupervised learning rule yields the result. Q.E.D.

Note that the second term of Eq(8a) satisfies the Hebbian product rule between output y and input x, but the first term computing the inverse matrix $|w|^{-1}$ is not scalable with increasing N nodes. This non-Hebbian learning enters through the logarithmic derivative of Jacobian giving $|\frac{dy}{dx}|^{-1}$. To improve the computing speed, S. Amari et al. assumed the identity matrix $[\delta_{i,k}] = [W_{i,j}][W_{j,k}]^{-1}$ and multiplied it to the BS algorithm $\frac{dH}{dW_{i,j}}[\delta_{i,k}] = \{[\delta_{i,j}] - (2\vec{y} - 1)\vec{y}^T]\}[W_{i,j}]^{-1}$, where use is made of $y_i = [W_{i,\alpha}]x_\alpha$ to change the input x_j to the synaptic gap by its weighted output y_i. In information geometry, Amari et al. derived the natural gradient ascend BSA algorithm:

$$\frac{dH}{dW_{i,j}}[W_{i,j}] = \{[\delta_{i,j}] - (2\vec{y} - 1)\vec{y}^T]\}, \qquad (8b)$$

which is not in the direction of original $\frac{dH}{dW_{i,j}}$ and enjoys a faster update without computing the BS inverse .

Fast ICA: Erkki Oja began his ANN learning of nonlinear PCA for pattern recognition in 1982.

$$<\vec{\tilde{x}}\vec{\tilde{x}}^T>\hat{e} = \lambda\hat{e};$$

$$w' = w + \vec{x}\sigma(\vec{x}^T\vec{w}) \cong <\vec{x}\vec{x}^T>\vec{w}; \qquad (8c)$$

$$\frac{d\vec{w}}{dt} = <\vec{x}\vec{x}^T>\vec{w} \cong \sigma(\vec{x}^T\vec{w})\vec{x} \cong \frac{dK(u_i)}{du_i}\frac{\mathrm{d}u_i}{\mathrm{d}w_i} \equiv \mathrm{k}(\vec{x}^T\vec{w})\vec{x};$$

where Oja changed the unary logic to bipolar logic $v_i = \sigma(u_i) \approx u_i - \frac{2}{3}u_i^3 \cong \frac{dK(u_i)}{du_i}$; $u_i = w_{i,\alpha}x_\alpha$. It becomes similar to a Kurtosis slope, which suggested to Oja a new contrast function K. In other words, Taylor expansion of the normalization, Eq(8c) and set $|\vec{w}|^2 = 1$:

$$|\vec{w}'|^{-1} = [(\vec{w} + \epsilon\vec{x}k(\vec{w}^T\vec{x}))^T\left(\vec{w} + \epsilon\vec{x}k(\vec{w}^T\vec{x})\right)]^{-\frac{1}{2}}$$

$$= 1 - \frac{\epsilon}{2}k(\vec{w}^T\vec{x})(\vec{x}^T\vec{w} + \vec{w}^T\vec{x}) + O(\epsilon^2).$$

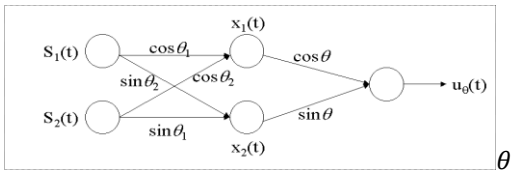
$$\vec{w}'' \equiv \vec{w}'|\vec{w}'|^{-1}$$

$$= \left(\vec{w} + \epsilon\vec{x}k(\vec{w}^T\vec{x})\right)\left(1 - \frac{\epsilon}{2}k(\vec{w}^T\vec{x})(\vec{x}^T\vec{w} + \vec{w}^T\vec{x})\right)$$

$$\Delta\vec{\boldsymbol{w}}'' = \vec{w}'' - \vec{w} = \boldsymbol{\epsilon}\left[\boldsymbol{\delta_{\alpha,\beta}} - \boldsymbol{w''_\alpha w''^T_\beta}\right]\vec{\boldsymbol{x}}_{\boldsymbol{\alpha}}\frac{dK(u_\beta)}{dw''_i} \qquad (8d)$$

This general derivation of BSA Eq(8b) is referred to by Szu collectively as **BSAO unsupervised learning**.

Fast ICA Example: A. Hyvarinen and Oja demonstrated Fast ICA in 1996, as the fixed point analytical solution: $\frac{dK(u_i)}{dw_i} = 0$, of a specific contrast function naned Kurtosis. Rather than maximizing an arbitrary contrast function, or the BS filtered output entropy, they considered the 4th order cummulant Kurtosis $K(y_i) = < y_i^4 > -3 < y_i^2 >^2$ which vanishes for a Gaussian average. K >0 for super-Gaussian, e.g. an image histogram that is broader than Gaussian, and K<0 for sub-Gaussian, e.g. a speech amplitude Laplacian histogram that is narrower than Gaussian. Every faces and voices have different fixed value of Kurtosis to set them apart. At the bottom of a fixed point, they set the slope of Kurtosis to zero and *efficiently* and *analytically solved* its cubic roots. This is called (Fast) ICA, as coined by Peter Como (Sig. Proc., circa '90).



$\theta \to \varphi$

$$\vec{x}_i = \vec{a}(\theta_i)_\alpha\vec{s}_\alpha = cos\theta_i s_1 + \sin\theta_i s_2; i = 1{,}2$$

$$\vec{u}_j = \vec{k}^T(\varphi_j)_\alpha\vec{x}_\alpha = cos\varphi_j x_1 + sin\varphi_j x_2; j = 1{,}2$$

$$\begin{pmatrix}u_1\\u_2\end{pmatrix} = \begin{bmatrix}\cos\varphi_1 & -\sin\varphi_1\\ \sin\varphi_2 & \cos\varphi_2\end{bmatrix}\begin{bmatrix}\cos\theta_1 & \sin\theta_2\\ -\sin\theta_1 & \cos\theta_2\end{bmatrix}\begin{pmatrix}s_1\\s_2\end{pmatrix}$$

$$= \begin{bmatrix}\cos(\varphi_1-\theta_1) & \sin(\varphi_1-\theta_2)\\ \sin(\varphi_2-\theta_1) & \cos(\varphi_2-\theta_2)\end{bmatrix}\begin{pmatrix}s_1\\s_2\end{pmatrix},$$

Oja rule of independent sources x & y:

$$K(ax+by) = a^4K(x) + b^4K(y)$$

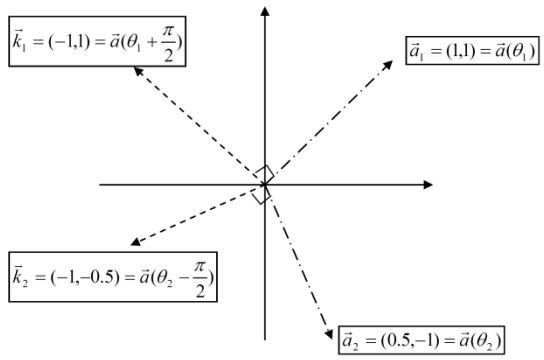
$$K(u_1) = \cos(\varphi_1-\theta_1)^4K(s_1) + \sin(\varphi_1-\theta_2)^4K(s_2)$$

$$K(u_2) = \sin(\varphi_2-\theta_1)^4K(s_1) + \cos(\varphi_2-\theta_2)^4K(s_2)$$

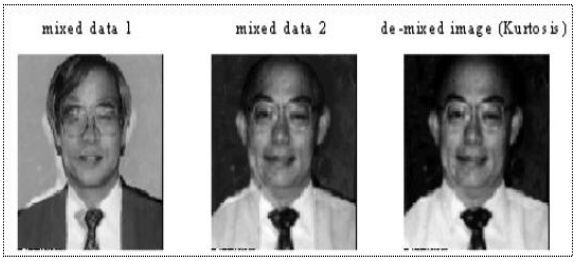
Given arbitrary unknown θ_i, not necessarily orthogonal to each other, the killing weight $\vec{k}(\varphi_j)$ can eliminate a mixing vector $\vec{a}_i(\theta_i)$. Szu's rule: Iff $\varphi_j = \theta_i \pm \frac{\pi}{2}$; then $K(u_1) = K(s_2); K(u_2) = K(s_1)$, verifing Fast ICA $\frac{\partial K}{\partial k_j} = 0$.

2.12 Sparse ICA

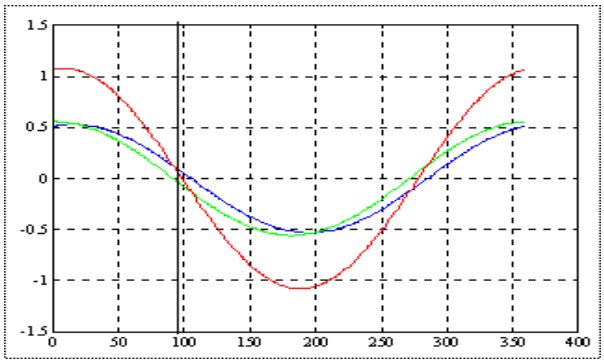
New application is applying a sparse constraint of non-negative matrix factorization (NMF), which is useful for image learning of parts: eyes, noses, mouths, (D. D. Lee and H. S. Seung, Nature 401(6755):788–791, 1999); following a sparse neural code for natural images (B. A.



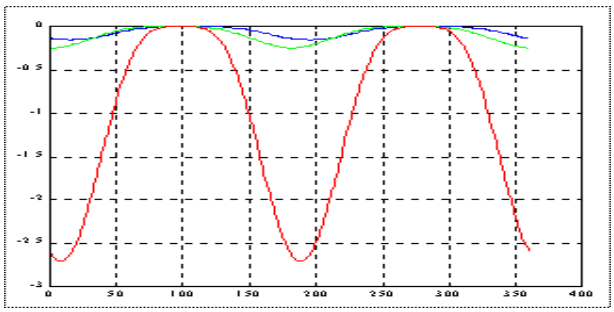
(a)



(b)



(c)



(d)

Figure 1: (a) De-mixing by killing vector(Yamakawa & Szu); (b) a sources image (Szu) (not shown Ymakawa) de-mixed by one of the killing vectors; (c)(left) The vertical axis indicates the blue source of Szu face vector, the green source of Yamakawa face vector, and the red is the Kurtosis value plotted against the killing weight vector. (d) (right) The Kurtosis is plotted against the source angle, where the max of Kurtosis happens at two source angles (Ref: H. Szu, C. Hsu, T. Yamakawa, "Adv. NN for Visual image Com," Int'l Conf Soft Computing, Iizuka, Japan, Oct. 1, 1998).

Olshausen and D. J. Field, Nature, 381:607–609, 1996). P. Hoyer provided Matlab code to run sparse NMF $[X] = [A][S]$, (2004).
$min.||[A]||_1; min.||[S]||_1$ subject to $E = |[X] - [A][S]|_2^2$.

The projection operator is derived from the Grand-Schmidt decomposition $\vec{B} = \vec{B}_\perp + \vec{B}_\parallel$; where $\vec{B}_\parallel \equiv (\vec{B} \star \vec{A})\vec{A}/|\vec{A}|^2$, and $\vec{B}_\perp \equiv \vec{B} - \vec{B}_\parallel$.

Alternative gradient descend solutions between 2 unknown matrices $\{[A] or [S]\}$:
new $[Z]' = [Z] - \frac{\partial E(|[X]-[A][S]|^2)}{\partial [Z]} = o;$ $min.||[Z]||_1$, where alternatively substituting $[Z]\ with\ [A]\ or\ [S]\}$

Recently, T-W. Lee and Soo-Young Lee, et al. at KAIST have solved the source permutation challenge of ICA speech sources in the Fourier domain by de-mixing for Officemate automation. They grouped similar Fourier components into a linear combination in a vector unit, and reduced the number of independent vectors in the sense of sparse measurements solving the vector dependent component analysis (DCA) [11].

2.13 Effortless Learning Equilibrium Algorithm

An effortless thought process that emulates how the e-Brain intuitive idea process works. Such an effortless thinking may possibly reproduce an intuitive solution, which belong to the local isothermal equilibrium at brain's temperature $K_B T_o$ (K_B is Boltzmann constant, $T_o = 273 + 37^o C = 310^o$ Kelvin). Therefore, the thermodynamic physics gives an inverse solution that must satisfy the minimum Helmholtz free energy: $min. \mathrm{E}(s_i) = E - T_o S$. The unknown internal brain energy is consistently determined by the Lagrange multiplier methodology. Thus, we call our m-component 'min-energy max a-priori source entropy' as Lagrange Constraint Neural Network (LCNN) in 1997. Taylor expansion of the internal energy introduced Lagrange parameter μ as variable energy slope.

$$E = E_o^* + \sum_{i=1}^{m} \frac{\partial E}{\partial s_i}(s_i - s_i^*) + O(\Delta^2) = E_o^* + \vec{\mu} \cdot ([W]\vec{X} - \vec{S}^*) + O(\Delta^2)$$

The Lagrange slope variable $\vec{\mu}$ were parallel and proportional to the error itself $\vec{\mu} \approx ([W]\vec{X} - \vec{S})$, our LCNN is reduced to Wiener LMS supervised learning $E \cong E_o^* + |[W]\vec{X} - \vec{S}^*|^2$ of the expected output $\vec{S}^*$ from the actual output $[W]\vec{X}$. Given the Boltzmann entropy formula: $S = -K_B \sum_i^k s_i \log s_i$, of independent s_i sources, the m-components general ANN formulism requires matrix algebra not shown here [9]. In order to appreciate the possibility of blind sources separation (BSS) of individual pixel, we prove the exact solution of LCNN for 2 independent sources per pixel as follows.

Exact Solution of LCNN: Theorem The analytical solution of LCNN of two sources is

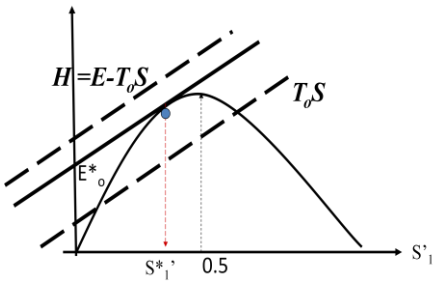


Figure 2: Exact LCNN pixel by pixel Solution

$$s_1^* = 1 - \exp(-\frac{E_o^*}{K_B T_o})$$

Derivation: Convert discret Boltzmann-Shannon entrpy to a single variable s_1 by normalization $s_1 + s_2 = 1$.

$$\frac{S(s_1)}{K_B} = -s_1 \log s_1 - s_2 \log s_2$$

$$= -s_1 \log s_1 - (1 - s_1) \log(1 - s_1),$$

We consider the fixed point solution:

$$Min.H = E - T_o S = 0;$$

so that

$$E = T_o\, S = -K_B T_o [s_1 \log s_1 + (1 - s_1) \log(1 - s_1)]$$

The linear vector geometry predicts another equation:

$$E = intercept + slope\ s_1 = E_o^* + \frac{dE}{ds_1}(s_1 - 0)$$

Consequently, the fixed point slope is computed

$$\frac{dE}{ds_1} = T_o \frac{dS}{ds_1} = T_o K_B (\log(1 - s_1) - \log s_1).$$

$$E = E_o^* + T_o \frac{dS}{ds_1} s_1 = E_o^* + T_o K_B (\log(1 - s_1) - \log s_1) s_1$$

Two formulas must be equal to each other at $s_1 = s_1^*$ yields

$$\frac{E_o^*}{K_B T_o} = -\log(1 - s_1^*). \quad \text{Q.E.D.}$$

The convergence proof of LCNN has been given by Dr. Miao's thesis using Nonlinear LCNN based on Kuhn-Tucker augmented Lagrange methodology. (IEEE IP V.16, pp1008-1021, 2007).

2.14 Interdisciplinary Contributions

Besides the aforementioned, the author is aware of the interdisciplinary contributions made by mathematicians (e.g. V. Cherkassky, Jose Principe, Lei Xu; Asim Roy); physicists (e.g. John Taylor, David Brown, Lyon Cooper); and biologists (e.g. Ishikawa Masumi, Mitsuo Kawato, Rolf Eckmiller, Shio Usui); as well as engineers (e.g. Kunihico Fukushima, K.S. Narendra, Robert Hecht-Nielsen, Bart Kosko, George Lendaris, Nikola Kassabov, Jacez Zurada, Cheng-Yuan Liou Of Taiwan U, You-Shu Wu of Tsing Hwa U, Huiseng Chi of Peking U, Toshio Fukuda, Hideo Aiso of 5th Gen Computing, et al.). The author apologized that he could not cover theirs and others contributions in this short survey.

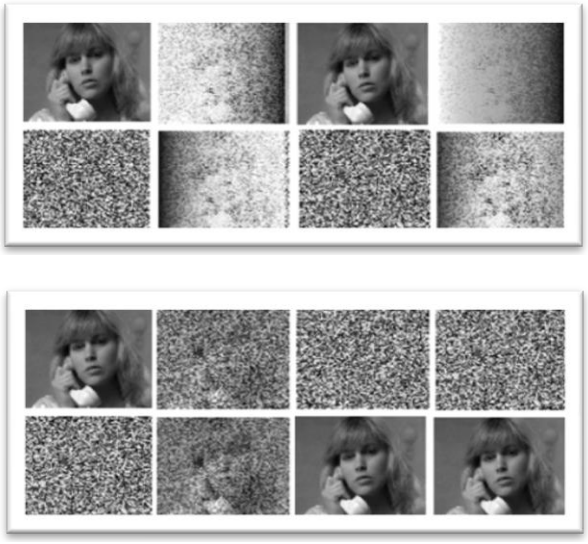


Figure 3: A face picture and a normal noise are mixed by a point nonlinearly (top panel) or linearly (bottom panel). The top panel has furthermore a column-wise changing mixing matrix, while the bottom panel has a uniform or identical mixing matrix. Since our LCNN is pixel-based for parallel implementation, it can solve both the top and the bottom panel (3rd column). However, the BSAO info-max algorithm is based on a batch ensemble working for a linear and identical mixing matrix (bottom panel 4th column only).

Combining ICA and CS linear algebras produced the Feature Organized Sparseness (FOS) theorem in Sect. 7. Contrary to purely random sparseness, the bio-inspired turns out to have the additional meaning, i.e., the locations indicating dramatic moment of changes at salient spatial pixel features. The measure of significance is quantified by the orthogonal degree among admissible states of AM storage [12]. Thus, the author reviewed only the math leading to ICA unsupervised learning to enlighten the Compressive Sensing (CS) community. Traditional ANNs learning are based on pairs of exemplars with desired outputs in the LMS errors, l_2-norm performance. A decade later, modern ANN has evolved close to the full potential of unsupervised connectionists (but still lack the architecture of self-organizing capability). We emphasize in this review that the HVS is a real time processing at a replenishing firing rate of about 1/17 seconds; HVS operates a smart AM tracking of those selected images in order to be retrieved instantly by those orthogonal salient features stored in the Hippocampus.

3. Neuroscience Organized Sparseness

We wish to help interdisciplinary readership appreciate the organization principle of spatiotemporal sparse orthogonal representation for Compressive Sensing. The purpose of sparse orthogonal representation is help increase the chance of pattern hits. A sparse orthogonal representation in

computer science is $\{e_i\} = \{(1,0,0,0..),(0,1,0,0,..),.;\ i = 1,2,.)\}$ known as a finite state machine. Human Visual System (HVS) has taken Hubel Wiesel oriented edge map wavelet $[\Psi_1,..]$ that transform a sparse orthogonal basis to another a sparse feature representation,

$$[\vec{f}_1,..] = [\Psi_1,..]^T[\vec{e}_1,...]$$

which is not a mathematically "Purely Random Sparseness," rather a biologically "Feature-orthogonal Organized Sparseness (FOS)". We shall briefly review Neuroscience 101 of HVS.

3.1 Human Visual Systems (HVS)

Physiologically speaking, the HVS has a uniformly distributed fovea composed of 4 million RG color vision cones for high resolution spot size. 2 millions B cones are distributed in the peripheral outside the central fovea in order to receive the high blue sky and the low blue lake water at 0.4μ wavelengths. This is understood by a simple geometrical ray inversion of our orbital lens when the optical axis is mainly focused on the horizon of green forest and bushes.

Based on Einstein's wavelength-specific photoelectric effect, the G cones, which have Rhodopsin pigment molecules sensitive to green wavelengths shorter than 0.5μ, can perceive trees, grasses, and bushes. Some primates whose G cone's Rhodopsin suffered DNA mutation of (M, L)-genes, developed a remarkable capability of spotting ripe red fruits hidden among green bushes with higher fructose content at a longer wavelength of about 0.7μ. These primates could feed more offspring, and more offsprings inherited the same trait, who then had more offspring, and so on, so forth. Abundant offspring eventually became tri-color Homo sapiens (whose natural intelligence may be endowed by God). Owing to the mutations, it is not surprising that the retina is examined by means of RGB functional stained florescence.

Millions of RG cones were found under a microscope arranged in a seemingly *random sparse* pattern among housekeeping glial (Muller) cells and B cones in the peripheral of the fovea. What is the biological mechanism for organizing sparse representation? If too many of them directly and insatiately send their responses to the brain whenever activated by the incoming light, the brain will be saturated, habituated, and complacent. That's perhaps why HVS developed a summing layer consisting of millions of Ganglions (gang of lions). Massive photo-sensors cones are located at the second layer, shielded behind a somewhat translucent ganglion layer. The Ganglions act as gatekeeper traffic cops between the eyes and the Cortex. A ganglion fires only if the synaptic gap membrane potentials surpass a certain threshold. This synaptic junction gap impedance can serve as a threshold suppressing random thermal fluctuations. It then fires in the Amplitude Modulation mode of membrane potential for a short distance, or Frequency Modulation mode of firing rates for a long distance.

3.2 Novelty Detection

Our ancestors paid attention to novelty detection defined by orthogonal property among local center of gravity changes. Otherwise, our visual cortex may become complacent from too many routine stimuli. Our ancestors further demanded a simple and rapid explanation of observed phenomena with paramount consequences (coded in AM of e-Brain). Thus, we developed a paranoid bias toward unknown events. For example, miracles must have messages, rather than accidents; or 'rustling bushes must be a crouching tiger about to jump out,' rather than 'blowing winds'. Thus, this meaning interpretation has been hard wired and stored in the AM of Hippocampus of e-brain. That's why in biomedical experiments, care must be given whenever rounding off decimals. A double-blind protocol (to the analyst and volunteer participants) with a (negative) control is often demanded, in order to suppress the bias of [AM] interpretation toward False Positive Rate. "In God we trust, all the rest show data." NIH Motto. We now know even given the data set, it's not enough unless there is a sufficient sampling in a double-blind with a control protocol (blind to the patients and the researchers who gets the drugs or the placebo, mixed with a control of no disease). Dr. Naoyuki Nakao thought in his e-brain about how to avoid potential kidney failure for high blood pressure patients who loseproteins. He could have been advocating an intuitive thought about the dual therapy of hypertension drugs; that both ACE inhibitor upon a certain hormone and ARB acting in a different way on the same hormone should be cooperatively administrated together. The paper appeared in the Lancet J. and since Jan. 2003, became the top 2 cited index in a decade. Swiss Dr. Regina Kurz discovered, "the data is too perfect to be true for small sample size of 366 patients." As a result, this dual drug therapy has affected 140K patients, causing a Tsunami of paper retractions at a 7 folds increase.

3.3 Single Photon Detection

The photons, when detected by cones or rods made of multiple stacks of disks, converts the hydrocarbon chain of Rhodopsin pigment molecules from the *cis* configuration to *trans* configuration. As a result, the alternating single and double carbon bonds of trans carbon chains are switching continuously in a domino effect until it reaches and polarizes the surface membrane potential. Then, the top disk has a 'trans state' and will not be recovered until it is taken care of at the mirror reflection layer, and converted back to the 'cis state' upward from the cone or rod base.

A single signal photon at the physiological temperature can be seen at night. No camera can do that without cryogenic cooling (except semiconductor Carbon Nano-Tube (CNT) IR sensor, Xi Ning & H. Szu). To detect a single moonlight photon, we must combat against thermal fluctuations at $300^o\ K \cong \frac{1}{40} eV$. How the thermal noise is cancelled without cooling operated at physiology temperatures. This is accomplished by synaptic gap junctions of a single ganglion integrating 100 rods. These 100 Rod bundles can

sense a single moonlight photon ($1\mu\sim1\ eV$) because there exist a 'dark light' current when there is no light, as discovered by Hagin [4]. Our eye supplies the electric current energy necessary to generate the 'dark currents.' It is an ion current made of Potassium inside the Rod and Sodium outside the Rod, circulating around each Rods. (i) Nature separates the signal processing energy from the signal information energy, because (ii) a single night vision photon does not have enough energy to drive the signal current to the back of brain; but may be enough (iii) to de-polarize the membrane potential to switch off the no signal 'dark current,' by 'negate the converse' logic. Any rod of the bundle of 100 rods receives a single photon that can change the rod's membrane potential to detour the 'Hagins dark currents' away from the Rod. Consequently, it changes the ganglion pre-synaptic junction membrane potential. As a result, the incoming photon changes the membrane potential and the ganglion fires at 100Hz using different reservoir energy budget for reporting the information [4]. A single ganglion synaptic junction gap integrating over these 100 rods bundle provides a larger size of the bundle to overcomes (v) the spatial uncertainty principle of a single photon wave mechanics. These (i~v) are lesson learned from biosensors. Another biosensor lesson is MPD computing by the architecture as follows.

3.4 Scale Invariance by Architecture

The pupil size has nothing to do with the architecture of the rod density distribution. The density drops off outside the fovea, along the polar radial direction in an exponential fashion. Thereby, the peripheral night vision can achieve a graceful degradation of imaging object size. This fan-in architecture allows the HVS to achieve scale invariance mathematically, as follows. These 1.4 millions night vision ganglion axon fibers are squeezed uniformly through the fovea channel, which closely packs them uniformly toward the LGN and visual cortex 17 in the back of head. The densities of Rods' and B-cones increase first and drop gently along the radial direction, in an exponential increase and decrease fashion:

$$Input\ locations = \exp(\pm Output\ uniform\ location),$$

which can therefore achieve a graceful degradation of the size variances by means of a mathematical logarithmic transformation in a MPD fashion without computing, just flow through with the fan-in architecture. This is because of the inverse $Output = \pm\log(Input) \cong \text{Output}'$, when Input = 2 x Input' because $\log(2)$ is negligible. This size invariance allows our ancestor to run in the moonlight while chasing after a significant other to integrate the intensity rapidly and continuously over the time without computational slow down. For photon-rich day vision, the high density fovea ganglions require 100 Hz firing rate, which might require a sharing of the common pool of resources, before replenishing because the molecular kinetics produces a natural supply delay. As a result, the ganglions who use up the resource will inhibit neighborhood ganglions firing rates, producing the lateral inhibition on-center-off-surround, the so-called Hubel and Wiesel oriented edge wavelet feature map $[\psi_n]$.[5]

3.5 Division of Labor

It's natural to divide our large brain into left and right hemispheres corresponding to our symmetric body limbs reversely. Neurophysiologic speaking, we shall divide our 'learning/MPD storing/thinking' process into a balanced slow and fast process. In fact, Nobel Laureate Prof. Daniel Kahneman wrote about the decision making by slow and fast thinking in his recent book published in 2011. We may explain the quick thinking in terms of intuitive thinking of the emotional side of right hemisphere (in short 'e-Brain') & the logical slow thinking at the left hemisphere, '*l*-Brain'. In fact, Eckhard Hess conducted experiments demonstrating pupil dynamics (as the window of brains) which is relaxed in a dilation state during a hard mental task which uses up mental energy and contracted iris to fit the intensity needed once the computing task is complete. We wish to differentiate by designing different tasks which part of the brain (*l*-brain, *e*-brain) is doing the task. This way we may find the true time scale of each hemisphere. For example, putting together a jigsaw puzzle depicting a picture of your mother or a boring geometry pattern may involve the *e*-Brain or *l*-Brain. How fast can our e-brain or *l*-brain do the job? In the cortex center, there are pairs of MPD storages called the hippocampus, which are closer to each other in female than male.

The female might be more advanced than male for a better lateralization and environmental-stress survivability. The faster learning of speech, when a female is young or the female has a better chance of recovery when one side of the brain was injured. Such a division of labors connected by the lateralization seems to be natural balance to build in ourselves as a self-correction mechanism.

3.6 Lateralization between e-Brain & *l*-Brain

According to F. Crick & C. Koch in 2005, the consciousness layer is a wide & thin layer, called Claustrum, located underneath the center brain and above the lower part lizard brain. The Claustrum acts like a music conductor of brain sensory orchestra, tuning at a certain C note for all sensory instruments (by the winner-take-all masking effect).The existence of a conscious toning remains to be experimentally confirmed (e.g. studying an anesthesia awakening might be good idea). It could be above the normal EEG brain waves types known as alpha, beta, theta, etc., and underneath the decision making neuron firing rate waves at 100 Hz. This pair of hippocampus requires the connection mediated by the Claustrum known as the Lateralization. According to the equilibrium minimum of thermodynamic Helmholtz free energy, the sensory processing indeed happens effortlessly at the balance between minimum energy and maximum entropy, we are operating at.

The sparse orthogonal is necessary for HVS, but also natural for brain neuronal representation. We have 10 billion neurons and 100 billion synapses with some

replenishment and regeneration, the synapses could last over 125 years. Another reason the sparse orthogonal representation is not loaded up with all the degree of freedoms and no longer has a free will for generalization. In other words, unlike a robot having a limited memory capacity and computing capability, we prefer to keep our brain degrees of freedom as sparse as possible, about 10 ~15% level (so-called the least developed place on the Earth) about 10% $x10^{20} \supseteq$ *encyclopedia Brittainca*. Todd and Marols in Nature 2004 [6] summarized the capacity limit of visual short-term memory in human Posterior Parietal Cortex (PPC) where sparsely arranged neuronal population called grandmother neurons fires intensely for 1 second without disturbing others, supporting our independence concept yielding our orthogonality attribute. The 'grandmother neuron(s)' may be activated by other stimulus and memories, but is the *sole* representation of 'grandmother' for the individual. To substantiate the *electric brain response* as a *differential response* of visual *event related potentials*, Pazo-Alvarez et al in Neuroscience 2004 [7] reviewed various modalities of brain imaging methodologies, and confirmed the biological base of feature organized sparseness (FOS) to be based on automatic *comparison–selection* of changes. "How many views or frames does a monkey need in order to tell a good zookeeper from a bad one?" Monkeys select 3 distinctive views, which we refer to as *m frames*: frontal, side and a 45^o view [8]. Interestingly, humans need only $m = 2$ views when constructing a 3-D building from architectural blueprints, or for visualizing a human head. These kind of questions, posed by Tom Poggio et al. in 2003 [8], can be related to an important medical imaging application.

4. Orthognal Sparse States of Associative Memory

Since the semiconductor storage technology has become inexpensive or 'silicon dirt cheap,' we can apparently afford wasteful 2-D MPD AM storage for 1-D vectors. Here, we illustrate how MPD AM can replace a current digital disk drive storage, a-pigeon-a-hole, without suffering recall confusion and search delays. The necessary and sufficient condition of such AM admissible states requires that rank-1 vector outer product is orthogonal as depicted in Fig.4. Thus, we recapitulate the essential attributes, sparseness and orthogonality as follows.

4.1 Connectionist Storage

Given facial images $\vec{X}_{N,t}$, three possible significant or salient features such as the *eyes*, *nose*, and *mouth* can be extracted in the rounding-off cool limit with the maximum firing rate of 100 Hz to one and lower firing rates to zero: (*1, 0*) ≡ (*big, small*). When these neuronal firing rates broadcast among themselves, they form the *Hippocampus* [AM] at the synaptic gap junctions denoted by the weight matrix $W_{i,j}$. For an example, when a small child is first introduced to his/her Aunt and Uncle, in fact the image of Uncle gets compared with Aunt employing the 5 senses. Further, fusion of information from all senses is conducted beneath the cortical layer through the Claustrum[13]. The child can distinguish Uncle by multi-sensing and noticing that he has a normal sized mouth (0), a bigger (1) nose as compare to Aunt, and normal sized eyes (0). These features can be expressed as firing rates $f_{old} \equiv (n_1, n_2, n_3) \equiv$ (*eye, nose, mouth*) ≡ (*0, 1, 0)* which turns out to be the coordinate $\hat{y}$ axis of the family feature space. Likewise, the perception of an Aunt with big (1) eyes, smaller (0) nose, and smaller (0) mouth (1,0,0) forms another coordinate axis $\hat{x}$. Mathematically k/N=0.3 selection of sparse saliency features satisfies the orthogonality criterion for ANN classifier. This ANN sparse classifier not only satisfies the nearest neighbor classifier principle, but also the Fisher's Mini-Max classifier criterion for intra-class minimum spread and inter-class maximum separation [9].
Alternatively, when Uncle smiles, the child generates a new input feature set $f_{new} \equiv (n_1, n_2, n_3) \equiv$ (*eye, nose, mouth*) ≡ (*0, 1, 1*) through the same neural pathway. Then the response arrive at the hippocampus where the AM system recognizes and/or corrects the new input back to the most likely match, the big-nose Uncle state (*0, 1, 0*) with a fault tolerance of direction $\cos(45^o)$. We write 'data' to the AM by an outer-product operation between the Uncle's feature vector in both column and row forms and overwrite Aunt's data to the same 2-D storage without cross talk confusion. This MPD happens among hundred thousand neurons in a local unit. The child reads Uncle's smile as a new input. The AM matrix vector inner product represents three feature neurons (0,1,1) that are sent at 100 Hz firing rates through the AM architecture of Fig. 1c. Further, the output (0,1,0) is obtained after applying a sigmoid σ_o threshold to each neuron which confirms that he remains to be the big nose Uncle.

4.2 Write

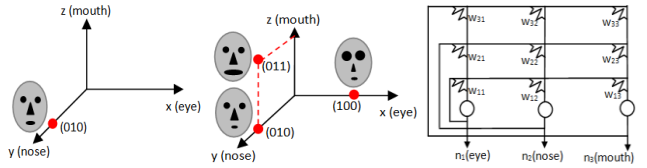
Write by the vector outer product repeatedly over-written onto the identical storage space forming associative matrix memory [AM]. Orthogonal features are necessarily for soft failure indicated in a 3-dimensional feature subspace of N-D.

$$[AM]_{big\ nose\ uncle} = \overrightarrow{output} \otimes \overrightarrow{input} = \begin{bmatrix}0\\1\\0\end{bmatrix} \begin{bmatrix}0 & 1 & 0\\0 & 0 & 0\\0 & 1 & 0\\0 & 0 & 0\end{bmatrix}$$

$$[AM]_{big\ eye\ aunt} = \overrightarrow{output} \otimes \overrightarrow{input} = \begin{bmatrix}1\\0\\0\end{bmatrix} \begin{bmatrix}1 & 0 & 0\\1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

MPD over-writing storage:

$$[AM]_{big\ nose\ uncle} + [AM]_{big\ eye\ aunt} = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 0\end{bmatrix}$$



(a) (b) (c)

Figur 4a: Feature organized sparseness (FOS) may serve as the fault tolerance attribute of a distributive associative memory matrix. When a child meets his/her Aunt and Uncle for the first time, the child pays attention to extract three features neurons which fire at 100 Hz or less, represented by 1 or 0 respectively. Figure 4b: Even if uncle smiled at the child (indicated by (0,1,1) in the first quadrant), at the next point in time, the child can still recognize him by the vector inner product read procedure of the [AM] matrix and the new input vector (0, 1, 1). A smiling uncle is still the uncle as it should be. Mathematically speaking, the brain's Hippocampus storage has generalized the feature vector (0, 1, 0) to (0, 1, 1) for a smiling big nose uncle, at the [AM] matrix. However, if the feature vector is over-learned by another person (0, 0, 1), the degree of freedom is no longer sparse and is saturated. In this case, one can no longer have the NI capability of the innate generalization within the subspace. Fig.4c: This broadcasting communication circuitry network is called the artificial neural network (ANN) indicating adjustable or learnable weight values {W_{ij}; i,j=1,2,3} of the neuronal synaptic gaps among three neurons indicated by nine adjustable resistances where both uncle and aunt features memory are concurrently stored.

4.3 Read

Read by the vector inner product recalling from the sparse memory template and employing the nearest neighbor to correct input data via the vector inner product:

$$Recall\ Vector\ =\ [AM][error\ transmitted] \cong$$

$$\sigma_o\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}0\\1\\0\end{bmatrix}\right)=\begin{bmatrix}0\\1\\0\end{bmatrix}\ smiling\ uncle\ remains\ to\ be\ uncle$$

4.4 Fault Tolerant

AM erases the one-bit error (the lowest bit) recovering the original state which is equivalent to a semantic generalization: a big nosed smiling uncle is still the same big nose uncle. Thus for storage purpose, the orthogonality can produce either fault tolerance or a generalization as two sides of the same coin according to the orthogonal or independent feature vectors. In other words, despite his smile, the AM self corrected the *soft failure degree about the degrees of sparseness* $30\% \cong \frac{k}{N} = 0.3$, or *generalized* the original uncle feature set depending on Claustrum fusion layer [13] for supervision. We demonstrate the necessary and sufficient conditions of admissible AM states that are sampled by the ***selective attention*** called the *novelty detection* defined by significant changes forming an ***orthogonal*** subspace. Further, the measure of significance is defined as degree of orthogonal within the subspace or not.

(i) We take a binary threshold of all these orthogonal novelty change vectors as the picture index vectors [12].
(ii) We take a sequential pair of picture index vectors forming a vector outer-product in the 2-D AM fashion.
(iii) Moreover, we take the outer product between the picture index vector and its original high resolution image vector in a hetero-associative memory (HAM) for instantaneous image recovery.

Thus, these 2-D AM & HAM matrix memory will be the MPD storage spaces where all orthogonal pair products are over-written and overlaid without the need of search and the confusion of orthogonal PI retrieval. Consequently, AM enjoys the generalization by discovering a new component of the degree of freedom, cf. Section 4.

5. Spatiotemporal Compressive Sensing

The software can take over tracking the local center of gravity (CG) changes of the chips- 1) seeded with the supervised associative memory of pairs of image foreground chip (automatically cut by a built-in system on chip (SOC)), and 2) its role play (by users in the beginning of videotaping). The vector CG changes frame by frame are accumulated to form a net vector of CG change. The tail of a current change vector is added to the head of the previous change vector until the net change vector becomes orthogonal to the previously stored net CG vector. Then, the code will update the new net CG change vector with the previous one in the outer product hetero-associative memory (HAM), known as Motion Organized Sparseness (MOS), or Feature-role Organized Sparseness (FOS). Then, an optical expert system (Szu, Caulfield, 1987) can be employed to fuse the interaction library (IL) matrix [HAM] (IL-HAM) in a massive parallel distributive (MPD) computing fashion. Building the time order [AM] of each FOS, MOS, and [HAM] of IL, we wish to condense by ICA unsupervised learning a composite picture of a simple storyline, e.g. YouTube/BBC on eagle hunting a rabbit.

We have defined [12] a significant event involving a local Center of Gravity (CG) movement such as tiger jumping out of fluffing around bushes (Fig.5). The processing window size may have a variable resolution with learnable window sizes in order to determine the optimum LCG movement. This may be estimated by a windowed Median filters (not Mean filter) used to select majority gray-value delegating-pixel locations and image weight values according to the local grey value histogram (64x64, 32x32, 16x16, etc.). Then we draw the optical flow vector from one local delegator pixel location to the next pointing from one to the next. The length of the vector is proportional to the delta change of gray values as the delegator weights. Similarly, we apply this Median Filter over all windows for two frames. We can sequentially update multiple frames employing optical flow vectors for testing the net summation. In this process, one vector tails-to-another vector head is plotted to cover a significant

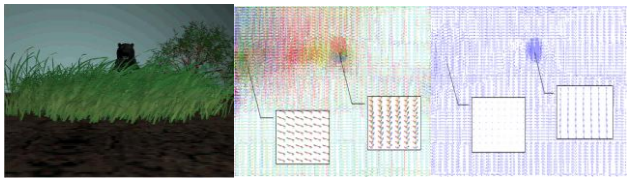


Figure 5: Net C.G. Automation [12]: (5a) Video images of a tiger is jumping out of wind-blowing bushes (Augmented Reality); (5b) The net Center of Gravity (CG) optical flow vectors accumulating f=5 frames reveals the orthogonal property to the previous net CG, capturing a Tiger jumping out region. This net CG motion of moving object (tiger) is different than wind-blowing region (bushes) in cyclic fluctuations; (5c) indicate an associated Picture Index sparse representation (dark blue dot consists of net CG vectors represented by ones, among short lines by zeros).

movement over half of the window size. Then the net is above threshold with the value of 'one' representing the whole window population to build a Picture Index (PI) (indicating a tiger might be jumping out with significant net CG movement); otherwise, the net CG will be threshold at zero (as the wind is blowing tree branches or bushes in a cyclic motion without a net CG motion). We could choose the largest jump CG among f frames.

Toward digital automation, we extracted the foreground from background by computing the local histogram based optical flow without tracking, in terms of a simplified medium filter finding a local center of gravity (CG). Furthermore, we generated the picture index (PI-AM) and the image-index (Image-HAM) MPD AM correlations [12]. We conjecture (TBD) another [HAM] of an interaction library (IL-AM) for fusion of storyline subroutine (Szu & Caulfield "Optical Expert System," App Opt. 1988) sketched as follows: AI pointer relational database, e.g. Lisp 1-D array (attribute (color), object (apple), values (red, green)) are represented by the vector outer products as 2-D AM maps. These maps are added with map frequency and restore a missing partial 2-D pattern as a new hypothesis. This type of interaction library can discover significant roles from selected foreground frames by generalizing AM. Further, this IL AM will follow the constructed storyline to compose these significant roles into a Video Cliff Note for tourist picture diary. For an example, a predator-prey video of about 4.5 minutes long was BBC copyrighted. Following the steps listed above, we have developed **compressive sampling (CSp)** video based on AM in terms of motion organized sparseness (MOS) as the picture index forming [AM] and its image as Hetero-AM [12]. Moreover, we have extended the concept with major changes shown as an automatic ***Video Image Cliff Notes.***

The lesson learned from the predator-prey BCC video is summarized in the Cliff Note, Fig.6, where a rapid change & stay-put as the keys for survivor(optical expert system, video Cliff Notes, SPIE DSS/ICA conf. Baltimore April, 2012).

6. Spatial-spectral Compressive Sensing Theory

We sketch a design of a new Smartphone camera that can take either daytime or nighttime picture with a single HVS focal plane array (FPA). Each pixel has a 2x2 Byres filter, which splits 1μ~1eV in quarter sizes and corrects different wavelength differences. Thus the filter trades off the spatial size resolution for increasing the spectral resolution. The camera adopts MPD [AM] & [HAM] storage in SSD medium. Such a handheld device may eventually become a personal secretary that can self-learn owner's habits, follow the itinerary with GPS during travel, and keep diary and send significant events.

We can relax the 'purely random' condition of the sparseness sampling matrix [Φ] with feature organized sparseness[Φ_s], where 1s indicate the locations of potential discovery of features. We shall derive a theorem to design

Figure 6: A predator-prey video of about 4.5 minutes long was taken from YouTube/BBC for education & research purposes. It emulated unmanned vehicle UXV (X=A,G,M) useful Intelligence Surveillance Reconnaissance. An eagle *cruising* gathered the intelligence by a few glimpse of a moving prey during the *surveillance* in the sky; after identifying it as a jumping rabbit, the eagle made a *chase engagement,* closed its wings and dropped at the rabbit. Rabbit was detected via moving shadow, stayed motionless avoiding motion detection. The rabbit jumped away from the ground zero whereas the eagle lost its ability to maneuver due to semi-closed wings at terminal velocity and suffered a heavy fall.

$[\Phi_s]$ by solving ICA unsupervised learning. Feature Organized Sparseness (FOS) Compressive Sensing works not only for video motion features, but also for color spectral features if we treat the spectral index as time index.

Theorem: Feature Organized Sparseness (FOS) Compressive Sensing:

ICA Unsupervised Learning methodology can help design FOS CS sampling matrix $[\Phi_s]$

$$[\Phi_s][\Psi] \equiv [ICA]; \quad [\Phi_s] = [ICA][\Psi]^{-1} \tag{9}$$

where $\{\psi_n\}$ is the Hubel-Wiesel wavelet modeled by the digital sub-band wavelet bases successfully applied to JPEG 2000 image compression and $\vec{s}$ is a column vector of feature sources

Proof: We can readily verify the result by comparing the CS linear algebra with the ICA unsupervised learning algebra side-by-side as follows:

$$R^N \ \vec{X} = \sum_{n=1}^{N} s_n \psi_n = \sum_{n_k=1}^{k} s_{n_k} \psi_{n_k} = [\Psi] \tag{10}$$

where k non-zeros wavelets are indexed by $n_k = 1,2,..,k \ll N$.

$$R^m: \ \vec{Y} = \sum_{i=1}^{m} x_i \phi_i^T = [\Phi_s]\vec{X} \ ; \tag{11}$$

Substituting Eq(7) into Eq(8), the linear matrix relationship yields a desired exemplar image $\vec{Y}$ which has the unknown mixing matrix [ICA] and the unknown feature sources $\vec{s}$

$$\vec{Y} = [\Phi_s][\Psi]\vec{s} \equiv [ICA]\vec{s}. \qquad Q.E.D.$$

We can exploit the full machinery of unsupervised learning ANN community about how to solve the Blind Sources Separation (BSS). We can either follow the Lagrange Constraint Neural Network based on minimizing the thermodynamic physics Helmholtz free energy by maximizing the a-priori source entropy [9] or the engineering filtering concept of maximizing the posterior de-mixed entropy of the output components [10]. For the edifice of the CS community that BSS is indeed possible, we have recapitulated the simplest possible linear algebra methodology with a proof as follow.

(i) *Symmetric Wiener Whitening in* ensemble average *matrix* $[W_z]^T = [W_z] =< \left[\vec{Y}\vec{Y}^T\right] >^{-\frac{1}{2}}$.

By definition $\vec{Y}' \equiv [W_z]\vec{Y}$ satisfying $< \vec{Y}'\vec{Y}'^T > \equiv [W_z] < \vec{Y}\vec{Y}^T > [W_z]^T = [I]$.

$$\because [W_z] < \vec{Y}\vec{Y}^T > [W_z]^T [W_z] = [I][W_z] = [W_z];$$

$$\therefore [W_z]^T [W_z] =< \left[\vec{Y}\vec{Y}^T\right] >^{-1}; [W_z] =< \left[\vec{Y}\vec{Y}^T\right] >^{-\frac{1}{2}} \quad \text{Q.E.D.}$$

(ii) Orthogonal Transform*:* $[W]^T = [W]^{-1}$

By definition

$[W]\vec{Y}' = [W][W_z]\vec{Y} = [W][W_z][\Phi_s][\Psi]\vec{s} \equiv [W][W_z][ICA]\vec{s} = \vec{s}$

$$\because [W] < \vec{Y}'\vec{Y}'^T > [W]^T \equiv [W][I][W]^T =< \vec{s}\vec{s}^T > \cong [I];$$

$$\therefore [W]^T = [W]^{-1} \qquad \text{Q.E.D.}$$

The Step (ii) can reduce ICA de-mixing to orthogonal rotation. We can compute from these desired exemplar images from their corresponding sources employing simple geometrical solutions called the killing vector. This vector is orthogonal to all row vectors except for one, cf. Fig. 1. Further the rotation procedure generates a corresponding independent source along the specific gradient direction.

Since we have applied (i) Wiener whitening in image domain, and (ii) orthogonal matching pursuit to derive the feature sources, we can pair the desired exemplar images with so-constructed feature sources by the rank-1 AM approximation of ICA mixing matrix [ICA]:

$$[ICA] = \sum \vec{y}\vec{s}^T = [\vec{y}_1, \vec{y}_2, \dots][\vec{s}_1, \vec{s}_2, ..]^T. \tag{12}$$

Our experience indicates a desirable orthogonality post-processing. Given all independent sources, we construct the orthogonal 1s (by the Gram-Schmidt procedure) $< \vec{s}\vec{s}^T > \cong [I]$.

$$[\Phi_s] \cong [\vec{y}_1, \vec{y}_2, \dots][\vec{s}_1, \vec{s}_2, ..]^T[\Psi]^{-1}. \tag{13}$$

Furthermore, we prefer the orthogonal feature extraction $[\Phi_s]$ such that $< [\Phi_s] \ [\Phi_s]^T > \cong [I]$. In doing so, we can increase the efficiency of multi/hyper-spectral compressive sensing methodology helping "finding a needle in a haystack" by sampling only the image correlated to the needle sources $\{\vec{s}_1, \vec{s}_2, .\}$ without unnecessarily creating a haystack of data cube blindly.

7. Handheld Day-Night Smartphone Camera

Our goal is making a new handheld smartphone camera which can take both daytime and nighttime pictures with a single photon detector array. It can automatically keepi and send only those significant frames capable of discovering motions and features. Our design logic is simple: never imaging daytime pictures with nighttime spectral, and vice versa, in a photon poor lighting or in the night do not take daytime color spectral picture. Of course, a simple clock time will do the job; but a smarter approach is through the correlation between exemplar images and desired features. We wish to design the camera with over-written 2-D storage in a MPD fashion, in terms of a FOS following the AM FT Principle. We can avoid the cross-talk confusion and unnecessary random access memory (RAM) search-delay, based on the traditional 1-D sequential optical CD technology storage concept: a pigeon-a-hole. This is a natural application of our Feature Organized Sparseness. We can build a full EOIR spectrum fovea camera applying a generalized Bayer filters using spectral-blind Photon Detectors (PD) emulating cones and rods per pixels. We mention that a current camera technology applied the Bayer color image filters (for RGB colors). We trade the spatial resolution with spectrum resolution. We take the spectral blind photon detector array of N pixels to measure N/4 color pixels. We modify the Bayer filter to be 4x4 per pixel and the extra 4th one is for extra night vision at near infrared 1 micron spectrum. We further correct optical path difference at new Bayer filter media in order to focus all spectrum on the same FPA, without the need of expensive achromatic correction in a compound lens.

Our mathematical basis is derived by combining both CS and ICA formulism, Eqs(6,7,8), and applying ICA unsupervised learning steps (i) & (ii) to design a FOS sampling matrix $[\Phi_s]$. Finding all the independent sources vectors from input day or night images $\vec{y}'s$ we collect expected sources $\vec{s}'s$ into a ICA mixing matrix $[ICA] = \sum \vec{y}\vec{s}^T$, then substituting its equivalence to CS sampling we can design FOS sampling matrix as $[\Phi_s] = [ICA][\Psi]^{-1}$ where $[\Psi]$ is usual image wavelet basis. The hardware is mapping the sparse feature sampling matrix onto 2x2 Bayer Filters per pixel that can afford to trade the spatial resolution with the spectral resolution in close up shots.

In this paper, we have further extended Motion Organized Sparseness [12] with Feature Organized Sparseness compositing two main players as a prey and a predator, namely a rabbit and an eagle. Their interaction is discovered by their chase after each other optical flows as shown in Fig. 6 as an automatic ***Video Image Cliff Notes***. Instead the purely random sparseness, we have generalized CS sampling matrix $[\Phi]$ with FOS sampling matrix$[\Phi_s]$.

In closing, we could estimate the complexity effect of replacing purely random sparseness $[\Phi]$ with FOS $[\Phi_s]$ upon the CRT&D RIP theorem. We could apply the complexity analysis tool called **Permutation Entropy** [14]. PE computes computed in a moving window of the size L=2,3, etc., counting the up-down shape feature of ones over the zeros: $H(L) \equiv -p(\pi)\sum p(\pi)$ of the *k*-organized sparseness to set a bound the sampling effect from purely random ones. For example, an organized sampling mask $[\Phi_s]_{m,N}$ had a row of {0,1,1,0,0,0…} which yielded a moving window of size L=2: in 4 cases[{01} up, {11}flat, {1,0}down; {0,0}flat, etc.]; size L=3 yields 3 cases[{0,1,1}up, {1,1,0}down; {1,0,0}down, {0,0,0} flat, etc.]. They had shown H(L) to be bounded from organized structure with ones locations (degree of complexity) to purely randomness (zero complexity) as $0 \le H(L) \le log\, L! \cong L\, logL - L$, by Sterling formula where $L \ll k \ll N$. $0 \cong PE([\Phi_s]_{m,N}) \ll PE([\Phi]_{m,N}) \cong 0(L)$ Therefore, instead of intractable l_0 -constraint, we could equally use l_1 -constrained LMS to both $[\Phi_s]_{m,N}$ and $[\Phi]_{m,N}$, if we were not already choosing HAM MPD for real time image recover [12].

Our teaching of the fittest survival may be necessary for early behaviors. The true survival of human species has to be co-evolved with other species and the environment we live in. This natural intelligence should be open and fair to all who are not so blindly focused by a narrowly defined discipline and ego. This imbalance leads to unnecessary greediness, affecting every aspect of our life. I publish this not for my need to survive; but to pay back the Communities who have taught me so much. The reader may carry on the unsupervised learning running on the fault tolerant and subspace-generalize-able connectionist architectures. Incidentally, Tai Chi practitioners by the walking meditation consider the Lao Tze's advocated mindless state, which is a balance between a fast thinking (Yin, light gravity weight) and a slow analyzing (Yang, heavy gravity weight). The transcendental meditation could achieve a low frequency brain wave (EEG Delta type), having the long wavelength reaching both sides of the hemispheres as the lateralization. If we were just relaxing your conscious-mind controlling muscles, and let the gravity potential takeover, the internal fluids that circulates freely insider our internal organs known as 'the Chi' can which can enhance the wellbeing about our physiology metabolism.

Acknowledgement

The author wishes to thank Dr. Charles Hsu and Army NVESD Mr. Jeff Jenkins, Ms. Lein Ma, Ms. Balvinder Kaur for their technical supports. Dr. Soo-Young Lee for his patience and encouragement. The author acknowledges US AFOSR grant support of CUA R/D as facilitated by Dean Prof. Charles C. Nguyen.

References

[1] E. J. Candes, J. Romberg, and T. Tao "Robust Uncertainty Principle: Exact Signal Reconstruction from Highly Incomplete Frequency Information," IEEE Trans IT, 52(2), pp.489-509, Feb.2006

[2] E. J. Candes, and T. Tao," Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies," IEEE Trans IT 52(12), pp.5406-5425, 2006.

[3] D. Donoho," Compressive Sensing," IEEE Trans IT, 52(4), pp. 1289-1306,2006

[4] W. A. Hagins (NIH), R. D. Penn, and S. Yoshikami; "Dark current and photocurrent in retinal rods," pp. 380-412, Biophys. J. VOL. 10, 1970; F. Rieke (U. Wash), D.A. Baylor (Stanford), "Single-photon detection by rod cells of the retina," Rev. Mod. Phys., pp. 1027-1036, Vol. 70, No. 3, July 1998.

[5] Hubel, D. H., and Wiesel, T. N. (1962). "Receptive fields and functional architecture in the cat's visual cortex." J. Neurosci, 160, 106–154.

[6] J. Jay Todd and Rene Marols, "Capacity limit of visual short-term memory in human posterior parietal cortex," Nature 428, pp.751-754, 15 Apr. 2004.

[7] P. Pazo-Alvarez, RE. Amenedo, and F. Cadaveira, "Automatic detection of motion direction changes in the human brain, " E. J. Neurosci. **19**, pp.1978-1986, 2004.

[8] Martin A. Giese and Tomaso Poggio, "Neural mechanisms for the recognition of biological movements," Nature Reviews Neuroscience 4, 179-192 (March 2003)

[9] H. Szu, IN: Handbook of Computational Intelligence IEEE Ch.16; H. Szu, L. Miao, H. Qi, "Thermodynamics free-energy Minimization for Unsupervised Fusion of Dual-color Infrared Breast Images," SPIE Proc. ICA etc. V. 6247, 2006; H. Szu, I. Kopriva, "Comparison of LCNN with Traditional ICA Methods," WCCI/IJCNN 2002; H. Szu, P. Chanyagorn, I. Kopriva: "Sparse coding blind source separation through Powerline," Neurocomputing 48(1-4): 1015-1020 (2002); H. Szu: Thermodynamics Energy for both Supervised and Unsupervised Learning Neural Nets at a Constant Temperature. Int. J. Neural Syst. 9(3): 175-186 (1999); H. Szu and C. Hsu, "Landsat spectral Unmixing à la superresolution of blind matrix inversion by constraint MaxEnt neural nets, *Proc. SPIE 3078*, pp.147-160, 1997; H. H. Szu and C. Hsu, "Blind de-mixing with unknown sources," *Proc. Int. Conf. Neural Networks*, vol. 4, pp. 2518-2523, Houston, June, 1997. (Szu et al. Single Pixel BSS Camera, US PTO Patent 7,355,182 Apr 8, 2008; US PTO Patent 7,366,564, Apr 29, 2008).

[10] A. J. Bell and T.J. Sejnowski, "A new learning algorithm for blind signal separation," adv. In Inf. Proc. Sys. **8**, MIT Press pp. 7547-763, 1996; A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," neu. Comp. 9, pp 1483-1492, July 1997; S. Amari, "Information Geometry," in Geo. and Nature Contemporary Math (ed. Nencka and Bourguignon) v.203, pp. 81-95, 1997.

[11] S. Y. Lee, et al. Colleagues, "Blind Sources Separation of speeches," and "NI Office Mate", cf. KAIST http://bsrc.kaist.ac.kr/new/future.htm.

[12] H. Szu, C. Hsu, J. Jenkins, J. Willey, J. Landa, "Capturing Significant Events with Neural Networks," to appear J. Neural Networks 2012

[13] F.C. Crick, C. Koch. 2005. "What is the function of the Claustrum?" Phil. Trans. of Royal Society B-Biological Sciences 360:1271-9.

[14] C. Brandt & B. Pompe, " Permutation Entropy-a natural complexity measure of time series,"PRL 2002; "Early detection of Alzheimer's onset with Permutation Entropy analysis of EEG," G. Morabito, A. Bramanti, D. Labate, F. La Foresta, F.C. Morabito, *Nat. Int.* (INNS) V.1, pp.30~32, Oct 2011.