A Compact VLSI Neural Computer Integrated with Active Pixel Sensor for Real-time Machine Vision Applications

Wai-Chi Fang, Suraphol Udomkesmalee, Leon Alkalai

Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, USA

ABSTRACT

A compact VLSI neural computer integrated with an active pixel sensor has been under development to mimic what is inherent in biological vision systems. This electronic eye-brain computer is targeted for real-time machine vision applications which require both high-bandwidth communication and high-performance computing for data sensing, synergy of multiple types of sensory information, feature extraction, target detection, target recognition, and control functions, The neural computer is based on a composite structure which combines Annealing Cellular Neural Network (ACNN) and Hierarchical Self-Organization Neural Network (HSONN). The ACNN architecture is a programmable and scalable multi-dimensional array of annealing neurons which are locally connected with their local neurons. Meanwhile, the HSONN adopts a hierarchical structure with nonlinear basis functions. The ACNN+HSONN neural computer is effectively designed to perform programmable functions for machine vision processing in all levels with its embedded host processor. It provides a two order-of-magnitude increase in computation power over the state-of-the-art microcomputer and DSP microelectronics. A compact current-mode VLSI design feasibility of the ACNN+HSONN neural computer is demonstrated by a 3-D 16x8x9-cube neural processor chip design in a 2-μm CMOS technology. Integration of this neural computer as one slice of a 4"x4" multichip module into the 3-dimensional MCM based avionics architecture for NASA's New Millennium Program is also described.

Keywords:

automatic target recognition, array processors, active pixel sensor, multichip module, neural networks, parallel processing, smart sensor, VLSI

1. INTRODUCTION

Machine vision is a challenge problem [1]. Many paradigms and algorithms have been proposed over the pass three decades toward this problem. However, a versatile solution to this problem has not emerged yet. The difficulty in obtaining a general and comprehensive solution is caused by the complexity of machine vision in itself. As shown in Figure 1, machine vision processing can be broken up into a hierarchy of low-level vision and high-level vision. Low-1evel processing must filter and segment an image into distinct regions and assign various characteristics (such as shape, color, texture, distance, and motion) to each region. High-level processing must recognize objects using information from low-level processing and provide target category assignment, reasoning, and decisions, The challenge of this problem is also caused by the enormous computation power required, the multiplicity of cues which must be detected and integrated, the difficulties of extracting 3D information from 2D image, variations in the sensing environment, limitations of the sensors, and noise. The success of a versatile machine vision system depends on succeeding at all these tasks.

Neural network approaches appear to be very promising for vision processing due to their massively parallel computing structures and learning capabilities, A number of studies have been reported on using artificial neural net works for machine vision applications [2, 3]. The existing works using neural networks share the same key idea that is performing the feature extraction or classification by defining an

object function using semi- or non-parametric methods to iteratively update the weights. The network with supervised learning is presented with both the input and the desired output for each input, and the weight structure best realizes the input/output relationship. The network with unsupervised learning is presented with the input data, and the weight structure is self-organized to follow data statistical regularities and group data into categories.

In this paper, a composite neural processor which combines Annealing Cellular Neural Network (ACNN) and Hierarchical Self-Organization Neural Network (HSONN) is proposed to provide a powerful neural computation engine for machine vision applications. Many ACNN+HSONN functions for vision processing and transformation have been verified via system simulation. These functions include noise filtering, isolated pixel elimination, hole filling, morphological operations, edge enhancement, edge detection, line detection, connected component detection, minimal and maximal detection, feature extraction, motion detection, motion estimation, etc. This composite neural structure can be effectively applied to processing at all levels of vision systems. Incorporating of the proposed VLSI neural processor into advanced vision systems offers orders-of-magnitude computing performance enhancements for on-board real-time processing and control tasks.

Section 2 introduces the ACNN+HSONN based neural computer named Eye-Brain Machine (EBM) which is a highly integrated multi-sensor and multi-processor system for vision applications. Section 3 describes the Annealing Cellular Neural Network and its associated VLSI neural processor design. Section 4 described the Hierarchical Self-Organization Neural Network and its associated VLSI neural processor design. Section 5 describes integration of this neural computer as one slice of a 10 cm x 10 cm multichip module into the 3-dimensional MCM based avionics architecture for NASA's New Millennium Program.
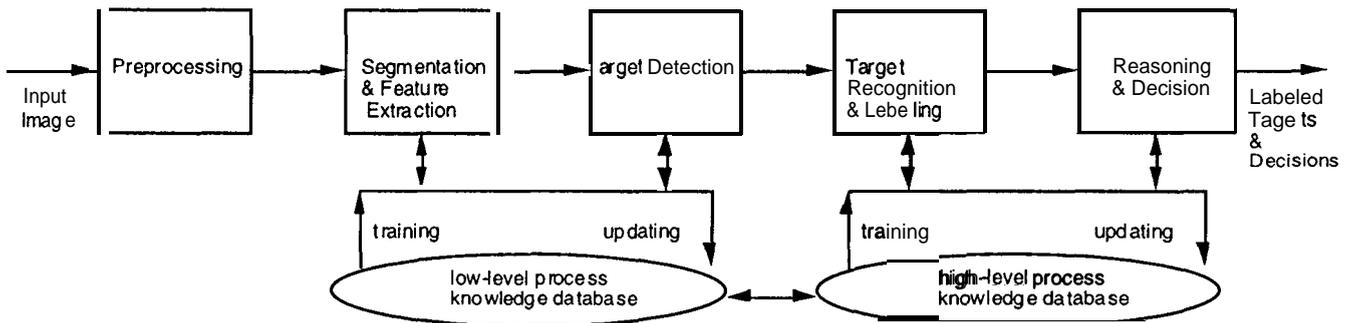

Figure 1: Functional block diagram of a general machine vision system.

2. AN EYE-BRAIN MACHINE FOR VISION APPLICATIONS

The ACNN+HSONN based neural computer named Eye-Brain Machine (EBM) has been under development at JPL. The Eye-Brain Machine is a highly integrated multi-sensor and neural processor system for vision applications. The design goal of the proposed EBM is to automatically recognize, localize, and classify point, area, and volume objects and phenomena in real-time at a 30 frame-per-second video rate.

The EBM includes two major subsystems: EBM-Eye, and EBM-Brain. The EBM Eye is a compact optoelectronic subsystem which integrates a wide range of different sensors with geometric, radiometric, and spectral parameters meeting the actual science and mission requirements. The EBM Brain is a high performance control and data subsystem which provides on-board computing resources for the EBM to perform various on-board tasks.

Figure 2 shows a functional design for the ACNN+HSONN neural computer. Major functional block of this neural computer includes multi-sensor parallel interface, parallel neural processor, on-line learning

coprocessor, digitally programmable synaptic weights, host system interface. A compact current-mode VLSI design feasibility of the neural processor was demonstrated by an 16x8x9-cube neuroprocessor chip design in a 2-μm CMOS technology. The whole neural computer can be accommodated into one slice of a 4"x4" multichip module.

For machine vision applications, the EBM neural computer can be used as a front-end sensory information processor to provide high throughput real-time computing power at neighborhood of the sensory circuit. Figure 3 shows a data flow diagram of the EBM design. An active pixel sensor [4] is integrated with the neural processor to accommodate the applications in which high-speed parallel external video inputs are needed, The integrated sensor and neuroprocessor architecture can take the combined advantages of the parallel image input and the parallel processing neural hardware to achieve ultra-high-speed sensory information processing.
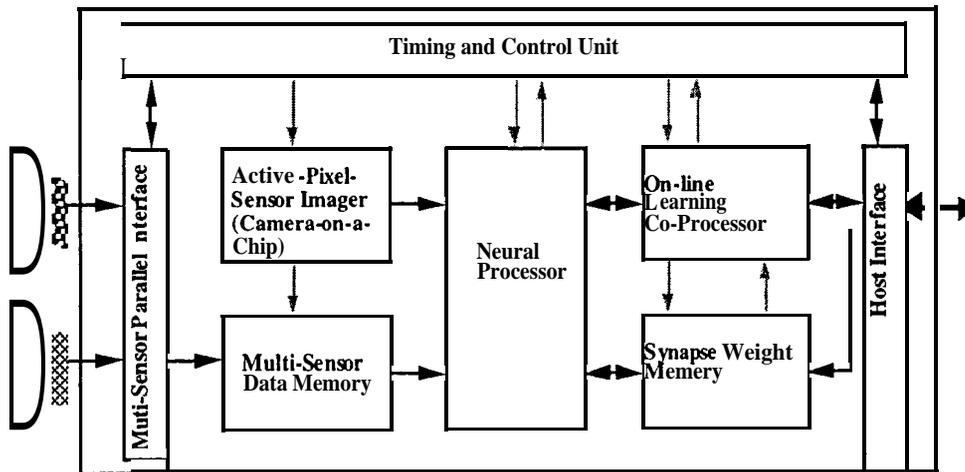

Figure 2. Functional Block Diagram for the ACNN+HSONN Neural Computer.
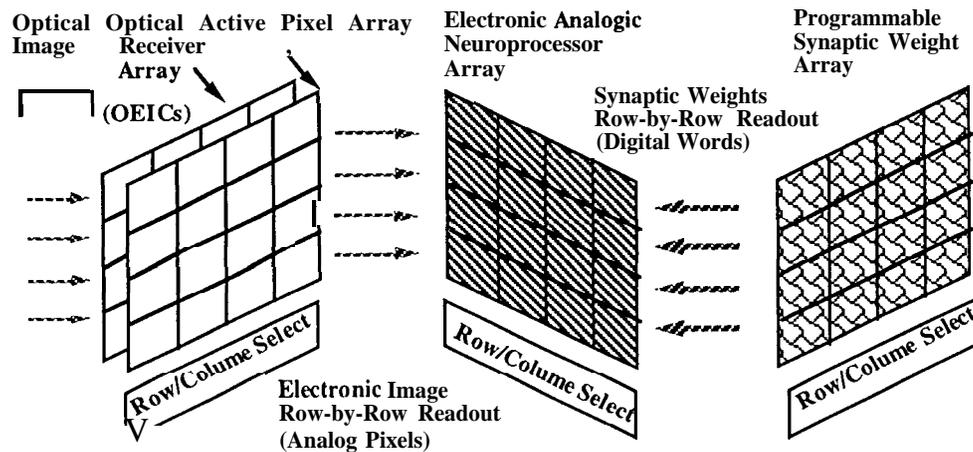

Figure 3. Data flow diagram of the integrated sensor and neuroprocessor.

3. ANNEALING CELLULAR NEURAL NETWORK

3.1. Introduction

Cellular Neural Networks is a multi-dimensional array of mainly identical cells which are dynamic systems with continuous state variables and locally connected with their local cells within a finite radius. Since its original publication by Chua and Yang [5,6] in 1988, the CNN paradigm has evolved rapidly and provides an unified framework for many computation-intensive applications such as signal processing and optimization. The CNN architecture is a locally connected, massively paralleled computing system with simple synaptic operators so that it is very suitable for VLSI implementation in real-time, high-speed applications. However, under the mild conditions [5], a CNN autonomously finds a stable solution for which the Lyapunov function of the network is locally minimized. To improve the local minimized energy function of the basic CNN, the ACNN is invented which use the hardware annealing method to accommodate the applications in which the optimal solutions of energy function are needed, Hardware annealing is a highly efficient method of finding optimal solutions for cellular neural networks [7], It is a paralleled version of fast mean-field annealing in analog networks. The ACNN is designed to perform programmable functions for fine-grained processing with annealing control to enhance the output quality. The digitally programmable synapse weights are designed for the ACNN to accommodate the applications in which programmable pre-determined operators are needed.

3.2.   Operation Theory

Consider a ACNN with n x m neurons as shown in Fig. 4.   Each neuron has the piecewise-linear transfer function $\&_w(.)$ and its gain is variable.   The gain is controlled by a monotonically increasing function $g(t)$ such that

$$v_y = f_{pw}(g\,v_x) = \begin{cases} +1 \text{ , if } v_x > \frac{1}{g} \\ g v_x \text{ ,if } -\frac{1}{g} \le v_x \le \frac{1}{g} \\ -1 \text{ , if } v_x < -\frac{1}{g} \end{cases} \tag{1}$$

The ACNN dynamic behavior satisfies a set of differential equations in the matrix notation as given:

$$C_x \frac{d\mathbf{x}}{dt} = -\frac{1}{R_x}\mathbf{x} + \mathbf{A}\mathbf{y} + \mathbf{B}\mathbf{u} + I_b\mathbf{w} \tag{2}$$

where $\mathbf{x} = [v_{x1} v_{x2} ... v_{xN}]^T$, $x = [v_{y1} v_{y2} ... v_{yN}]^T$, $\mathbf{u} = [v_{u1} v_{u2} ... v_{uN}]^T$, and $\mathbf{w} = [1\ 1...1]^T$ is an Nx1 unity vector, N= n x m. The capacitance C_x and resistance R_x at the state node of the processing element, and bias current I_b are assumed to be the same for the whole network. In (2), $\mathbf{A}$ and $\mathbf{B}$ are two-real N-by-N matrices consisting of feedback and feed-forward synapse weights and determined by given cloning templates T$_A$ and T$_B$, respectively. For the shift-invariant ACNNS, they are real symmetric. Since a piecewise-linear function is used as the transfer function of the network, the generalized energy function is a scalar-valued quadratic function of the output vector y,

$$E = -\frac{1}{2}\sum_{i,\,j}\sum_{C(k,l)\in N_r(i,j)} A(i,j;k,l)\ v_{yij} v_{ykl} + \frac{1}{2gR_x}\cdot\sum_{i,j}(v_{yij})^2 - \sum_{i,j}\sum_{C(k,l)\in N_r(i,j)} B(i,j;k,l) v_{yij} v_{ukl} - \sum_{i,j} I_b v_{yij}$$

$$= -\frac{1}{2}\mathbf{y}^T\left(\mathbf{A} - \frac{1}{gR_x}\mathbf{I}\right)\mathbf{y} - \mathbf{y}^T\mathbf{B}\mathbf{u} - I_b\mathbf{y}^T\mathbf{w} = -\frac{1}{2}\mathbf{y}^T\mathbf{M}_g\mathbf{y} - \mathbf{y}^T\mathbf{b}. \tag{3}$$

where $\mathbf{M}_g = \mathbf{A} - \frac{1}{g}\left(\frac{1}{gR_x}\right)\mathbf{I}$ and $\mathbf{b} = \mathbf{B}\mathbf{u} + I_b\mathbf{w}$. Notice that $T_x = \frac{1}{R_x}$

Here the factor l/g in the second term stems from the energy associated with the piecewise-linear function with a neuron gain other than unity.   By noting that $\mathbf{M}_g = \mathbf{A} - T_x\mathbf{I} - ((1-g)T_x/g)\mathbf{I} = \mathbf{M} - ((1-g)T_x/g)\mathbf{1}$, the relationship between the eigenvalues of un-annealed and annealed network can be easily shown to be

$$\lambda_k = \lambda'_k - \frac{(1-g)}{gR_x} = \lambda'_k - \frac{(1-g)T_x}{\pi g}, k = 1, 2, \ldots . N$$

(4)

where λ'_k are the eigenvalues of **M.** In the hardware annealing, the eigenvalues λ_k 's are changed from all negative initial values to the final values λ'_k 's by increasing the neuron gain g, such that the energy function (3) which is initially a convex function of y, is transformed gradually into a concave function. Regardless of initial state values, the network results in the optimal solution at which its energy is minimized globally,

3.3. 3-Dimensional ACNN Design for Vision Applications

In the original 2-D CNN every pixel is represented by one neuron. In the 3-dimensional ACNN every pixel can be represented by multiple neurons which form a hyperneuron and execute the maximum evolution function for various profile selections or data synergy. The 3-D ACNN designed for motion detection is shown in Figure 5. A set of *(2Dx + 1)(2Dy + 1)* layers of neurons are used to represent the optical flow field. Each layer corresponds to a different velocity and contains N_r x N_c neurons if the images are of size N_r x N_c. All neurons in the same layer are self-connected and locally interconnected with other neighboring neurons in a smoothing window of size *(2Wr + 1)(2Wc + 1)* . There are no interconnections between neurons in the different layers except those in the same hypercolumns. Every image pixel is represented by *(2Dx + 1)(2Dy + 1)* mutually exclusive neurons which form a *hyperneuron* for velocity selection. When the *(i, j, k, l)*-th neuron at the point *(i, j)* in the *(k, l)*-th layer is active high, the velocity of pixel *(i, j) is kB* and *lB* in the *k* and *l* direction, respectively. Here, *B is* the sampling bin size of the velocity component range. The interconnection weights $T_{i,j,k,l;m,n,k,l}$ consist Of the *smoothness constraints* and the *line process* only. The bias input $I_{i,j,k,l}$ contain all information from the images. And thus the error function for computing the optical flow is transformed into the *energy function* of the neural network that is defined by

$$E = -\frac{1}{2} \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=-D_x}^{D_x} \sum_{l=-D_y}^{D_y}$$

$$(\sum_{(m-i, n-j) \bullet SO} T_{i,j,k,l;m,n,k,l} v_{m,n,k,l} v_{i,j,k,l} + I_{i,j,k,l} v_{i,j,k,l}),$$

(5)

Optical flow computation is performed by the neuron evaluation using a massively parallel updating scheme which is based on the minimal mapping theory. Only the winning neuron is active high and the other neurons of the same hyperneuron are turned off. The network operation will be terminated whenever the energy function of the network reaches a minimum.

3.3. **VLSI Design for ACNN Neuroprocessor**

Building blocks and circuit designs for the ACNN neural processor include: digital programmable synapse, hardware annealing circuit, summation circuit, the nonlinear transfer function circuit, active pixel sensing circuit, etc. To construct a complete ACNN, a multiple of the units can be arranged in an n-by-m rectangular grid with appropriate interconnections. A 3-D ACNN can be realized in a time-multiplexed fashion or in a multiple ACNNS configuration.

To illustrate the implementation feasibility, a 16x8x9-cube ACNN chip of active dimensions 12.5 mm x11.7 mm was designed in a 2-µm CMOS technology. A programmable 5x5-array ACNN prototype chip was fabricated of active dimensions 1380 µm x 746 pm. A circuit board was built to demonstrate the operation of this prototype chip. Experiments on edge detection were performed. The measured result of

the edge detection experiment agrees well with the C-based simulation result. The CPU time for the C-Chased simulation is 2.53 seconds. The speedup is about 160,000.
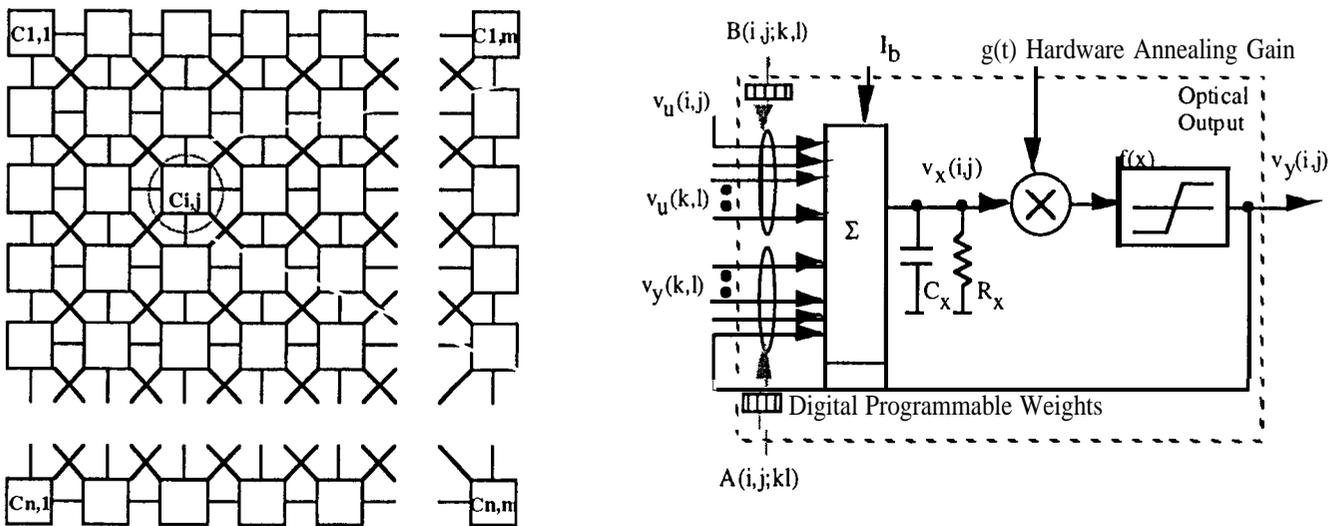

Figure 4. (a) An n-by-m ACNN array on rectangular grid. The shaded boxes are the neighborhood cells of C(i,j). (b) Model of the ACNN neuron C(i,j).
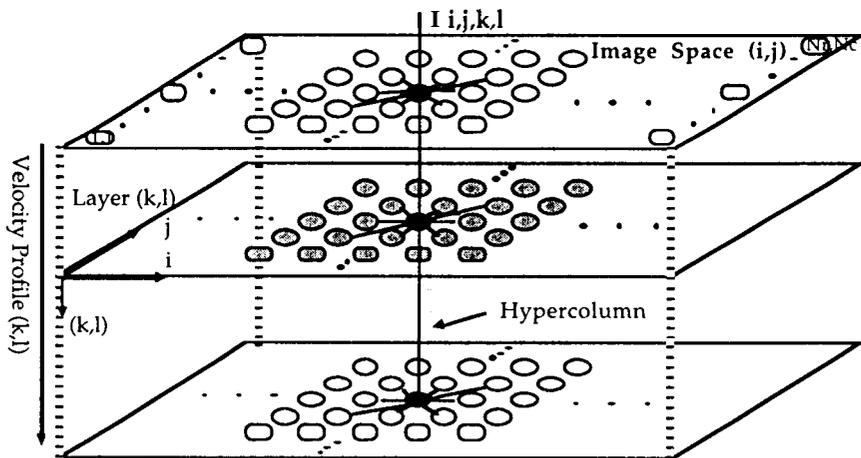

Figure 5. A 3-D cube multilayer ACNN model for motion flow computing.

4 Hierarchical Self-Organization Neural Network (HSONN)

4.1. Introduction

The success of most machine vision applications is closed tied with the reliability of the recognition of 3D objects or surfaces from 2D images. The hierarchical self-organization neural network (HSONN) is proposed to provide a high performance recognize. It effectively performs recognition to find a correspondence between certain features in the image and similar features of the object model.

The fundamental theory of self-organizing networks was presented by Grossberg [9], Kohonen [10], and other researchers [11]. To improve the learning efficiency and the recognition reliability, the HSONN is based on a hierarchical structure which is composed of a global winner-take-all layer and a set of local winner-take-all competitive networks.

4.2   HSONN Architecture and Learning Process

As shown in Figure 6, the global layer is a winner-take-all network with supervised learning. Meanwhile, the local layer is a set of local winner-take-all networks with self-organization learning. Each local winner-take-all network is trained by feature- or profile-specific data vectors for corresponding feature or profile selections. The local winner-take-all network systematically distributes the training data vectors in the vector space R^n to approximate the unknown probability density function of the training data vectors. The synapse weight vectors quantize the vector space and converge to cluster centroids.

For the global winner-take-all network, the supervised learning is used. The network is presented with both the input and the desired output for each input, and the weight structure best realizes the input/output relationship, Then the trained global layer serves as a classifier to encode the local winners. The global winner-take-all network is used to recognize the global winner from all local winners.
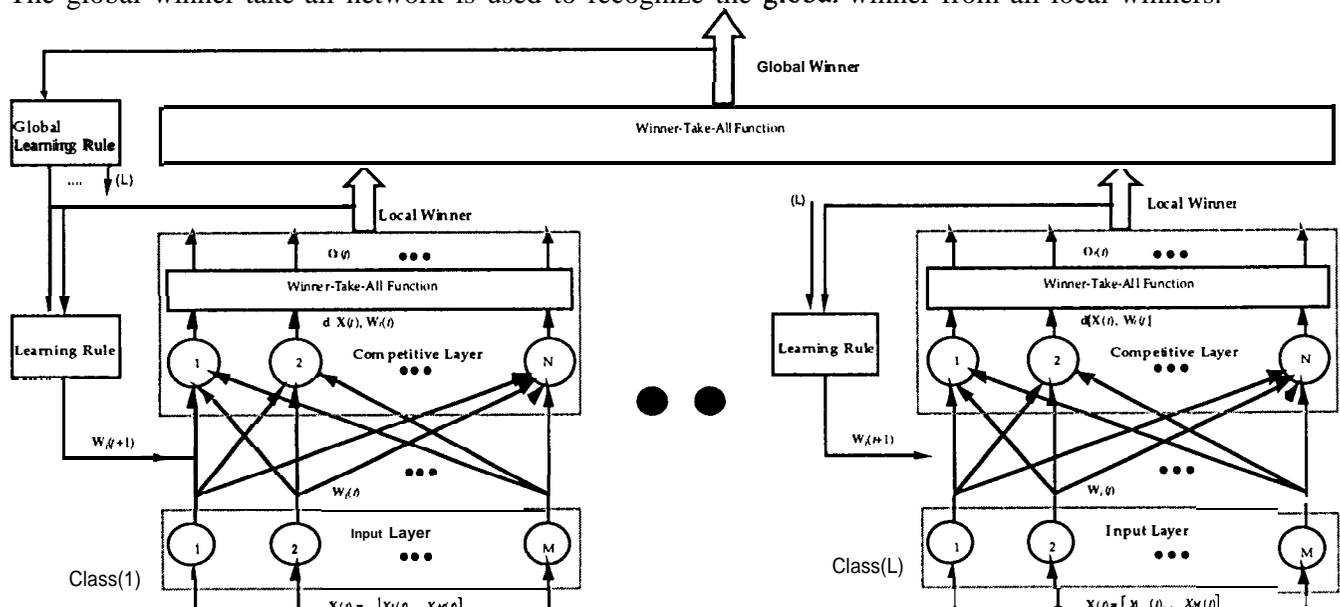

Figure 6. Hierarchical Self-Organization Neural Network.

4.3   VLSI Neural Processor Design

The VLSI architecture design for the local self-organization network is shown in Fig. 7. The M input neurons correspond to the elements of the M-dimensional input vector. Each input neuron gets its input from the external data bus and distributes the buffered signal to N distance-computing neural units in the competitive layer through the synapse matrix. Each distance-computing neuron calculates a square of Euclidean distance between its synapse weight vector and the input vector. The competitive process is performed throughout the whole layer by the winner-take-all operation. The winning neural unit is determined according to the minimum distance criterion. The synapse weights are then updated according to the local learning rule.

A mixed-signal VLSI design technique is used to implement the HSONN neural processor. The analog circuitry performs massively paralleled neural computation and digital circuitry processes multiple-bit address information. This neural-based recognize realizes a full-search vector quantization process for each input vector at a time complexity $O(1)$. It consists of the input neurons, programmable synapses, summing neurons, winner-take-all cells, and an index encoder. The programmable synapse matrix is composed of M x N synapse cells, which correspond to N M-dimensional synapse weight vectors. The output neuron array is composed of N summing neurons, which perform paralleled summation of the distortions between the input vectors and synapse weight vectors. The winner-take-all block consists of N competitive circuit cells which perform paralleled comparison among N *inverted* distortion values and choose a single winner. This block also provides a sufficiently high output level for the winning neuron against the rest. The index encoder circuit is an N-to-n decoder that uses binary codes to encode N classes. The on-line learning unit is a coprocessor to support a high-speed neural-network learning algorithm.
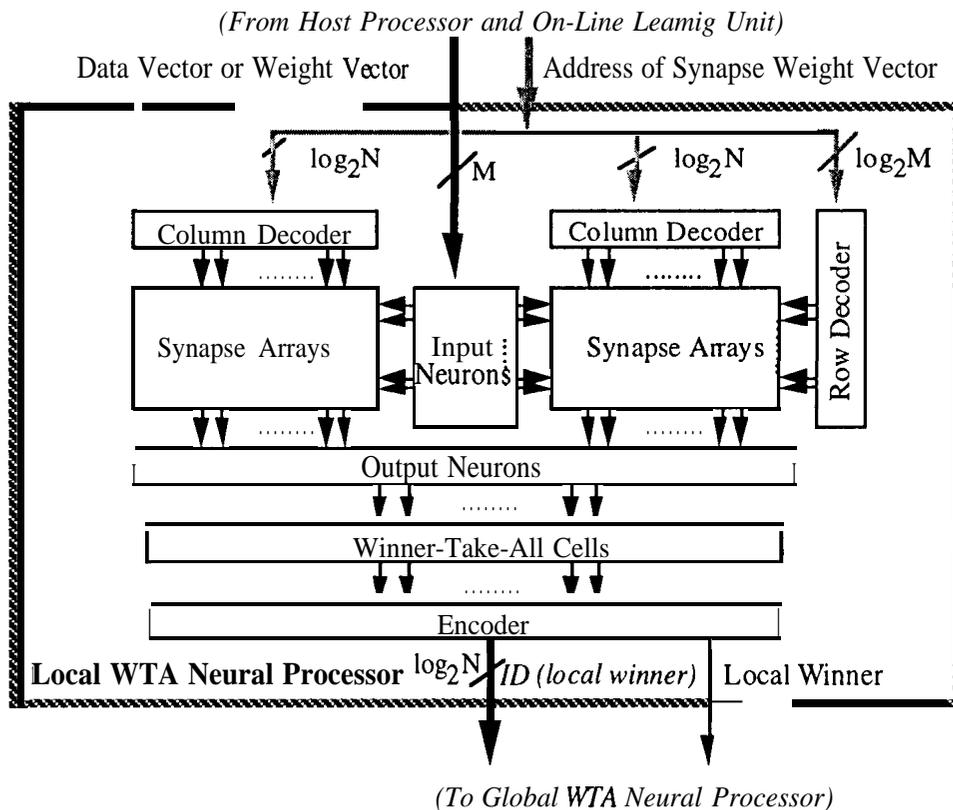
Figure 7. VLSI architecture design for the local self-organization neural network.

5. SYSTEM INTEGRATION WITH 3-D MCM BASED AVIONICS ARCHITECTURE

An ultra-low-power 3-D 1024x1024x27-cube ACNN neural processing engine has been under a development study by using a 3-D VLSI die stacking technology combined with a sub-0.25 µm low power ($V_{DD} = 0.9$ V) SOI CMOS process technology. A 3-D VLSI die stack of dimensions 4 cm x 4 cm x 1 cm is projected to accommodate a complete ACNN+HSONN processor with one 1024x1024 active-pixel sensor on the top of the die stack. The intrinsic computation power of the 3-D 1024x1024x27-cube ACNN neural computer is about up to 30 pets-connections per second at a 40-MHz evaluation rate. A miniaturized 30-pets-connection Eye-Brain Machine is therefore feasible to be implemented into a compact multichip module at a manageable power dissipation rate.

Integration of this neural computer as one slice of a 10 cm x 10 cm multichip module into the 3-dimensional MCM based avionics architecture for NASA's New Millennium Program is also feasible. Figure 8 shows the projected 3-D MCM stack for 4 multichip module designs: neural computer module, microcomputer and memory module, massive memory storage module, and communication and utility module.
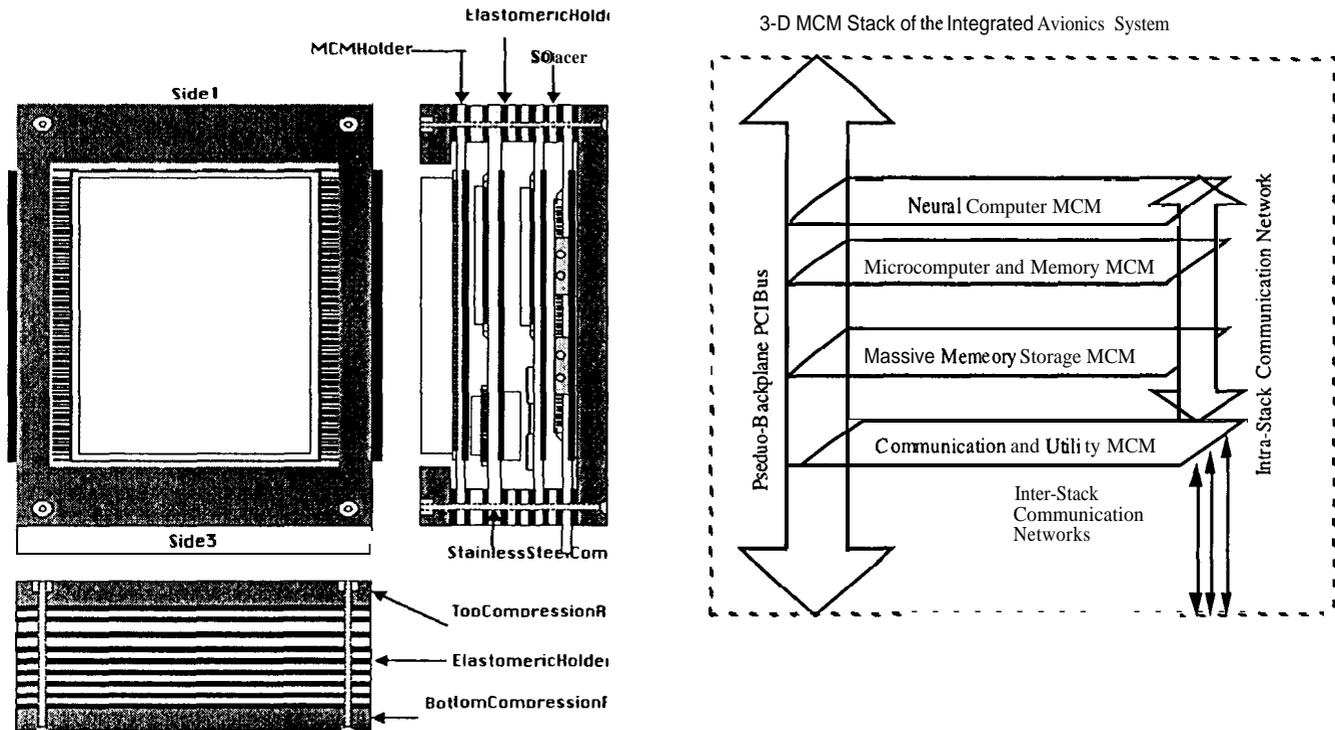

Figure 8. Integration of the neural computer MCM into the 3-D MCM based avionics architecture for NASA's New Millennium Program.

6. CONCLUSIONS

Machine vision is a challenge problem. The difficulty in obtaining a general and comprehensive solution is caused by the complexity of machine vision in itself. Neural network approaches appear to be very promising for vision processing due to their massively parallel computing structures and learning capabilities.

In this paper, we propose a neural network based machine vision system named Eye-Brain Machine (EBM) which is a highly integrated multi-sensor and multi-processor system for vision applications. The neural network is a combination of Annealing Cellular Neural Network and Hierarchical Self-Organization Neural Network. The ACNN architecture is a programmable and scalable multi-dimensional array of annealing neurons which are locally connected with their local neurons. The HSONN adopts a hierarchical structure with nonlinear basis functions.

The ACNN+HSONN neural computer is effectively designed to perform programmable functions for machine vision applications with its embedded host processor. VLSI neural processor design feasibility for the ACNN+ HSONN composite network is demonstrated. Incorporating of the proposed VLSI neural processor into advanced vision systems offers orders-of-magnitude computing performance enhancements for on-board real-time processing and control tasks. Major design features of this neural computer includes massively parallel neural processing, hardware annealing capability, winner-take-all mechanism,

digitally programmable synaptic weights, multi-sensor parallel interface, low-power VLSI design, and advanced multichip module packaging. Integration of this neural computer as one slice of a 10 cm x 10 cm multichip module into the 3-dimensional MCM based avionics architecture for NASA's New Millennium Program is also shown to be feasible.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] D.H. Ballard, C. M. Brown, *Computer Vision,* Prentice-Hall: Englewood Cliffs, N.J.

[2] S. Y. Kung, J.-N. Hwang, "Neural Network Architectures for Robotics Applications," IEEE Trans. on Robotics and Automation, Vol. 5, No. 5, Oct. 1989.

[3] B. Sheu, J. Choi, Neural Information Processing and VLSI, Kluwer Academic Publishers: MA, 1995.

[4] Fossum and P. Wong, "Future prospects for CMOS active pixel sensors," 1995 IEEE Workshop on CCDS and Advanced Image Sensors, Dana Point, CA, April 20-22, 1995.

[5] L. O. Chua, L. Yang, "Cellular neural networks: Theory", IEEE Trans. on Circuits and Systems, vol. 35, pp. 1257-1272, Oct. 1988.

[6] L. O. Chua, L. Yang, "Cellular neural networks: Applications," IEEE Trans. on Circuits and Systems, vol. 35, pp. 1273-1290, Oct. 1988.

[7] B. Sheu, S. Bang, W.-C, Fang, "Optimal solutions of selected cellular neural network applications by the hardware annealing method," Proceedings of the third IEEE International Workshop on Cellular Neural Networks and their Applications (CNNA-94), pp. 279-284.

[8] S. Grossberg, "Adaptive pattern classification and universal recording: I. Parallel development and coding of neural feature detectors, " *Biological Cybernetics,* vol. 23, pp. 121-134, *1976.*

[9] T. Kohonen, *Self-Organization and Associative Memory,* 2nd Ed., Spring-Verlag: New York, NY, 1988.

[10] R. Hecht-Nielsen, "Application of counterpropagation networks," *Neural Networks,* International Neural Network Society, vol. 1, pp. 131-141, 1988.

[11] L. Alkalai, J. Klein, M. Underwood. "The New Millennium Program M·icroelectronics Integrated Product Development Team Technology Roadmap", JPL D- 13276, November, 1995.