Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

Implementation of multi-layer leaky integrator networks on a cellular processor array

David R. W. Barr, Piotr Dudek, Jonathan M. Chambers and Kevin Gurney

Abstract - **We present an application of a massively parallel processor array VLSI circuit to the implementation of neural networks in complex architectural arrangements. The work was motivated by existing biologically plausible models of a set of sub-cortical nuclei - the basal ganglia. The model includes 5 layers, each consisting of 16384 leaky integrator neurons, with inter-layer synaptic weights forming various one-to-one and diffuse connectivity patterns. The architecture of the SIMD processor array allows all the neurons per layer to be updated simultaneously. The performance of the processor array chip in simulating the model is compared with the original model being executed on a computer workstation. It is demonstrated that in this application the chip outperforms the workstation by five orders of magnitude in terms of computational performance and seven orders of magnitude in terms of energy efficiency, providing a high-speed, low-power, compact hardware platform for possible embedded robotic applications.**

I. Introduction

Neural networks have proved to be useful in many applications from intelligent systems to pattern classification, but have the common characteristic that they are difficult to implement efficiently using standard processing architectures [1]. It is therefore justifiable to develop and use novel computational techniques and hardware architectures to solve this processing problem. In particular, the use of cellular processor arrays has been proposed [2, 3, 4] to improve computational efficiency and achieve real-time performance. This paper presents a technique for processing a certain class of biologically plausible neural networks, using the massively parallel processing architecture of the SCAMP-3 cellular processor array chip [5]. The implemented network is a 128x128 channel model of the basal ganglia (BG), a group of highly interconnected deep brain nuclei thought to play a role in action selection. This system complements a larger visual attention system by selecting between potential saccadic eye movements, using salience information from the retina and visual cortex to determine the most "visually interesting" location in space. It should be noted that the implemented network architecture was not chosen and crafted to showcase the processor array, but was determined by an existing biologically constrained model [8].

David R. W. Barr and Piotr Dudek are with the School of E & EE, University of Manchester, PO Box 88, M60 1QD, United Kingdom. (e-mail: d.barr@postgrad.manchester.ac.uk, p.dudek@manchester.ac.uk)

Jonathan M. Chambers, Kevin Gurney are with the Department of Psychology, University of Sheffield, Western Bank, Sheffield, S10 2TP, United Kingdom. (e-mail: j.m.chambers@sheffield.ac.uk, k.gurney@sheffield.ac.uk)

a) The Basal Ganglia Model

The basal ganglia are a group of sub-cortical nuclei, including striatum, subthalamic nucleus, globus pallidus and substantia nigra. They receive excitatory input from most regions of the cerebral cortex and send inhibitory outputs to multiple nuclei in the brainstem and thalamus. The basal ganglia are thought to be critical for solving the problem of action selection – the resolution of competition between command centres in the brain seeking behavioural expression in limited motor resources. Further, we suppose that the nuclei comprising the BG are connected topographically forming a series of parallel, repeated circuits or *channels*, representing individual actions [6].

An interpretation of the functional anatomy of the BG as a feed-forward on-centre, off-surround network acting as a signal selection mechanism has been proposed in [7]. Under this scheme signals input to the BG represent the *salience* of different action requests. The computation implemented by the BG circuitry acts to arbitrate between these by causing the normally tonic (quiescent) inhibitory output of the BG to pause in the most salient channel, while raising inhibitory output on losing channels. This model has been extended in [8] to explore the relationship between the BG and the oculomotor system, one of the best understood sensorimotor modalities. Much of the oculomotor system is retintotopically organised, that is, visual and motor activity relating to adjacent points in the visual field are processed by adjacent neural populations in a given nucleus. It was proposed that retinotopy is preserved within the oculomotor BG, and in the projections linking the BG to the oculomotor system.

The functional circuitry of the BG model is shown in Fig.1. Duplicate salience input from cortex/thalamus is sent to the sub-thalamic nucleus (STN) and striatum, which is further sub-divided in two groups of cells classified according to the way they utilise the neuromodulator dopamine (DA) via differing receptors (SD1 and SD2). The globus pallidus internal segment (GPi) and substantia nigra pars reticulata (SNr) - which together form the output nuclei of the BG - send inhibitory projections back to thalamus and to motor nuclei in the brainstem. Spontaneous, tonic activity in the STN guarantees that this output is active by default, so that

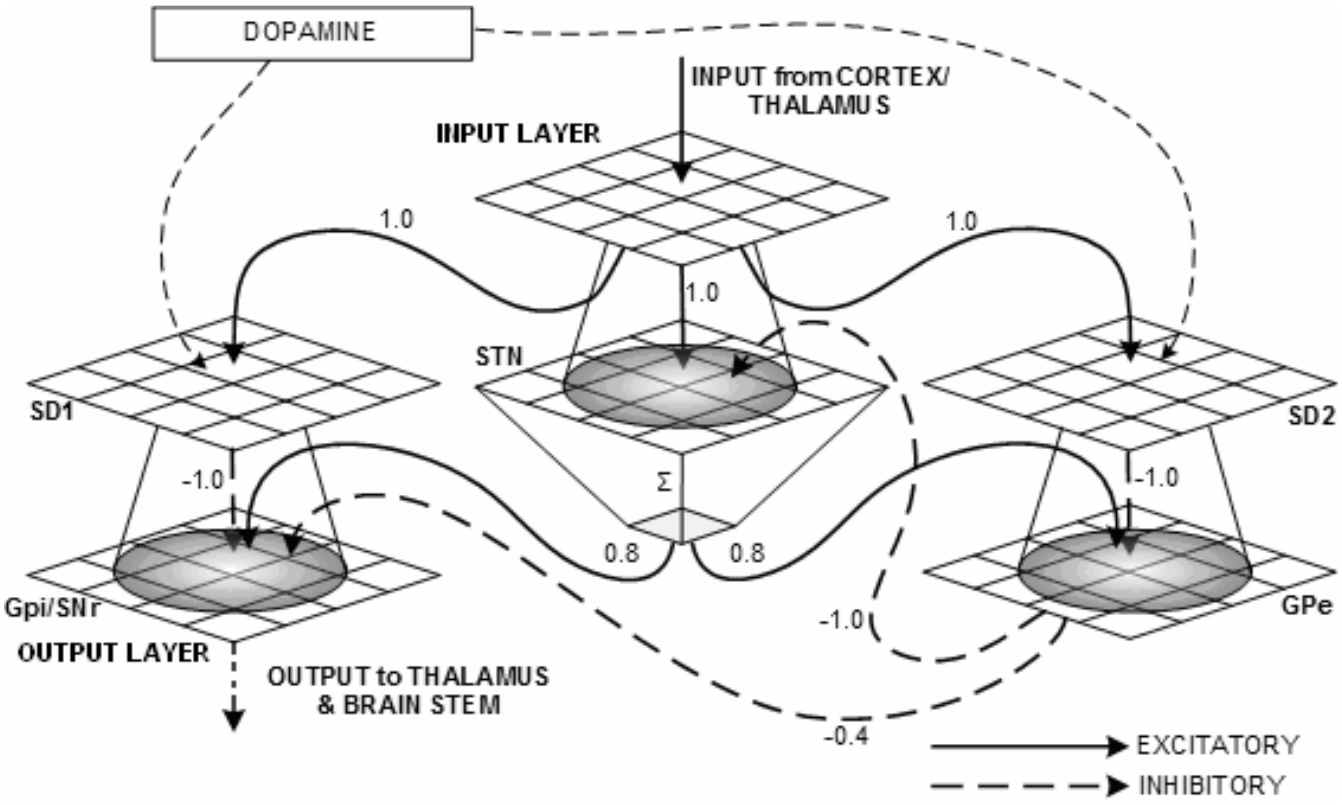


Figure 1 - Basal ganglia model (see text for details)

all motor systems are blocked. The extent to which a channel is selected is determined by the difference between its own activity and a measure of total activity in all channels. This calculation takes place in GPi/SNr, where diffuse excitatory input from the STN provides a measure of total activity in all channels, and focused inhibitory input from SD1 striatal cells provides a measure of individual channel activity. The excitation provided by STN is offset in the most active channel by the inhibitory input from SD1, causing a pause in the output of GPi/SNr for that channel. STN input goes unchecked in less active channels thus acting to block unwanted actions. This system has been described [7] as the *selection pathway*. The *control pathway* incorporates the globus pallidus external segment (GPe), and provides capacity-scaling by ensuring that STN activity does not become excessively high when multiple channels have non-zero salience, thus assuring full disinhibition of the winning channel irrespective of the number of competing channels. The model was implemented using rate coded model neurons - or 'leaky integrators' – in order to capture the key interactions between the nuclei without having to fix a raft of neuronal parameters.

The implemented oculomotor BG model [8] uses large arrays of neurons to represent the retinotopic topography. The connections between neural layers are either one-to-one (channel-wise), or use diffuse, but topographically localised projective fields with Gaussian weight profiles (Input to STN, SD1 to GPi/SNr and SD2 to GPe), as shown in Fig.1. By selecting the most salient channel the oculomotor BG are in effect choosing the most behaviourally significant point in visual space, and by producing a localised pause in inhibition delivered to oculomotor structures, ultimately leading to a gaze shift towards the region of interest.

b) The SCAMP-3 Vision Chip

The SCAMP-3 chip is an analogue cellular processor array. A more detailed description of this device can be found in [5], here we briefly outline its architecture. The chip comprises a 2D array of 128x128 processors, each with 9 registers and the ability to communicate with its direct neighbours, see Fig. 2. An instruction is broadcast to all the processors simultaneously, and each operates on its own local data, using the single instruction, multiple data (SIMD) paradigm. The spatial arrangement of the processors make the SCAMP-3 chip ideal for image processing through the use of integrated photo-detectors coupled with each processor. Therefore images can be focussed onto the array, sampled, operated upon, and then the output used in subsequent systems. As each pixel has its own processor, traditionally time-consuming data-parallel operations take a fraction of the time. Pixel values are stored as analogue currents in registers, and optimised computations are performed in the analogue domain. As there are nine registers per processor, the equivalent of nine 128x128 grey-level images can be stored on chip. Furthermore, a resistive-grid like structure embedded in the inter-processor communication network facilitates fast execution of diffusion operation. While the chip has been designed for image processing, its general-purpose processor array architecture allows efficient implementation of a class of

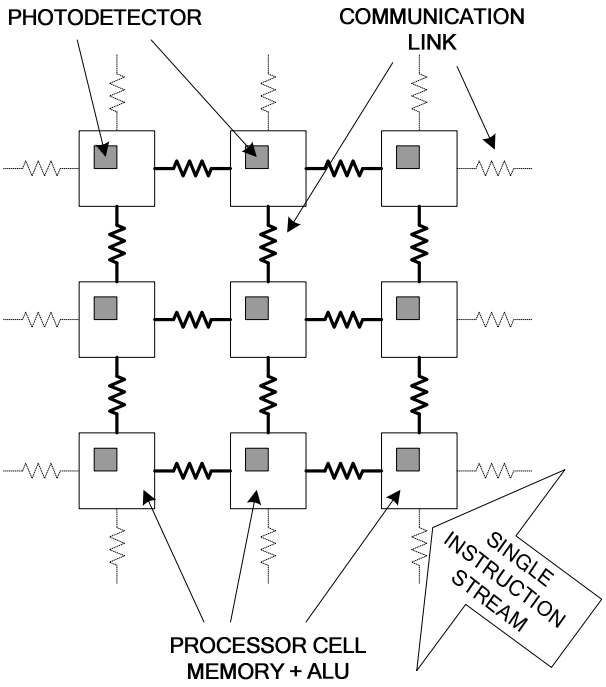


Figure 2 - A SCAMP-3 region of processor cells (each including photodetector, local memory and ALU) organized in a 4-neighbour mesh, with embedded resistive grid.

massively parallel systems, including multi-layer, topographically mapped (e.g. retinotopic) leaky integrator neural networks.

As the SCAMP-3 is a vision chip, integrating image sensing with the processor array, it is feasible that models of both the retina and a version of the oculormotor system can be implemented entirely in the SCAMP-3 architecture. This integration would take advantage of its low power, compact implementation and be useful in embedded biologically inspired autonomous agents. By moving the bulk of the homogenous low-level operations to the SCAMP-3, additional processing resources and bandwidth become available, allowing more sophisticated high-level models to be implemented on the host system.

II. Implementing The Model

Since the SCAMP-3 processor array maps naturally onto large scale models (with one processor per model node) we implemented 16384 channels (equal to the number of available processors) on a retinotopic grid of 128x128 cells.

a) Leaky Integrators

The neurons activities within a layer form a 2D array which can be stored in a SCAMP-3 array register. The neurons are implemented as leaky integrators. This means that only the activity level of the neuron need be stored The BG model has

```
D = I – A
for x = 1 to n
    D = D / 2
next x
A = A + D
```

Figure 3 - Leaky integrator neuron as implemented on the SCAMP-3 system. Registers D, I and A are arrays, x and n are scalars. A represents activities of all neurons in a layer, I is the corresponding pre-synaptic input layer. Arithmetic operations on arrays are performed in parallel, element-wise.

5 layers, requiring 5 registers within the processors. This leaves a further 4 registers to perform temporary calculations. As the SCAMP-3 chip operates on all of the pixels of an image at once, all of the neurons in a layer can be updated simultaneously.

A leaky integrator can be considered a temporal low-pass filter, implemented as:

$$a_{t+1} = a_t + \frac{\Delta t}{\tau}\left(i - a_t\right) \qquad (1)$$

where a is the neuron activity, i is the input calculated as a suitably weighted sum of pre-synaptic layers, Δt is the discrete simulation time-step and τ is the time-constant. Due to the specific limited arithmetic instruction set of the SCAMP-3 chip the division-by-two is a preferred operation, therefore the time coefficient has been implemented as:

$$\frac{\Delta t}{\tau} = \frac{1}{2^n} \qquad (2)$$

where n is an integer value. Thus the time constant can be adjusted in discrete time-steps, relative to the simulation time-step, as $\tau = 2^n \Delta t$ This provides sufficient accuracy and flexibility within the capabilities of the system. Pseudocode in Fig. 3 shows how the equations (1)-(2) have been mapped onto the SCAMP-3 architecture.

b) Output function

The output of each layer is simply the activation levels of the neurons within that layer. No explicit output function is applied, however due to the non-linear characteristic of the analogue storage cells, within the SCAMP-3 processors (shown in Fig. 4), a sigmoid output function is obtained.

c) Inter-layer connectivity

Consider each pixel within an image to be a neuron, the intensity of that pixel representing the output of the neuron. Connectivity and weighting between neuron layers can be distributed in certain patterns, by applying image processing

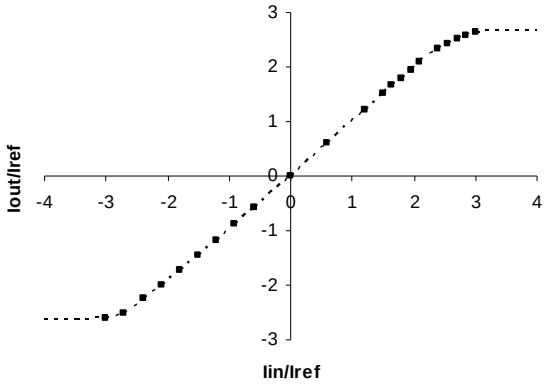


Figure 4 - The sigmoid like characteristic of the analogue storage cell on the SCAMP-3 chip. The graph shows measured results obtained after two inverting current-transfer operations [10]. Input current I_{in} and output current I_{out} are normalised to the reference current I_{REF}.

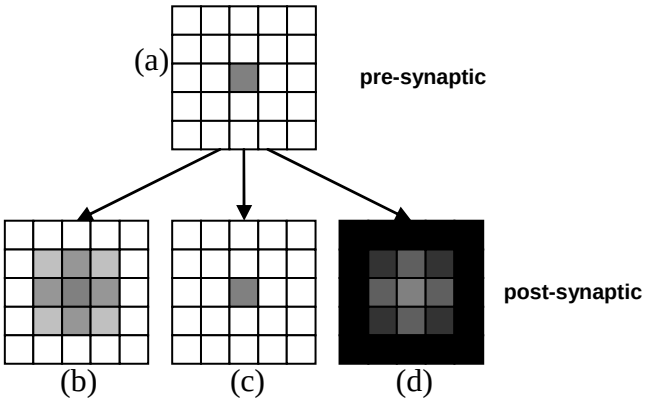


Figure 5 - Images can be transformed to represent connectivity patterns (a) pre-synaptic neuron output; (b) 3x3 Gaussian; (c) one-to-one; (d) Inhibitory 3x3 Gaussian.

transformations such as diffusion to the output of the neuron layer as shown in Fig. 5. In this way, projections from one layer to another (or within a layer) are determined by the transformation (e.g. convolution kernel). The complete neural input to a layer is calculated as a weighted summation of pre-synaptic output layers, after the relevant transformation. The SCAMP-3 chip excels at performing array-wide Gaussian projections in particular, as its spatial arrangement of processors and the analogue processing paradigm allows the near instantaneous Gaussian blurring of a particular register. The diffusion operation is implementated using a 'resistive grid' circuitry. This offers a significant reduction in processor time compared to that of a standard processing architecture.

In line with the large scale model of [8] layer-to-layer connectivity was determined by Gaussian receptive and projective fields, as well as the discrete one-to-one channel-wise connectivity. Some changes to the model were forced by implementation constraints of the SCAMP-3 chip, and these concerned the inter-layer connection weights. The SCAMP-3 chip has limited support for writing information to individual pixels, and does not support accurate multiplication between two registers. Fortunately, the homogeneous nature of the BG model's channels removes the need for individual neuron parameters and the model weights are close to values that are easy to implement in a SCAMP-3 environment, such as 0.5, 1, -1.

Only three weight changes to the original BG model [7] were required, to adapt it to SCAMP-3 architecture. The first was the synaptic weight from GPe to GPi, which has been changed from -0.4 to -0.5. The remaining two changes were a modification to the output of the STN, to both GPi and GPe, allowing for an increase in the number of channels. The original BG model scaled the STN output summation by 0.8, whereas the adapted model scales by 3.0/*n*, where *n* is the number of channels. Influence from dopamine has been omitted due to the difficulty in scaling registers. A solution to this could be to offset neuron input by a constant which is possible within the SCAMP-3 system. The output of the STN layer involves a summation across the entire array, which would usually be a costly operation, but the availability of a global array summation in the SCAMP-3 architecture makes this a simple task.

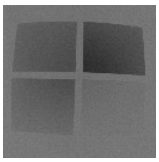
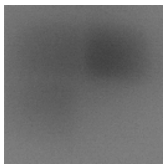
In the present implementation, the input to the BG model comes from the photo-detectors of the SCAMP-3 chip. Test images are generated and presented to the chip. In a more complete model, the visual input can be pre-processed by the SCAMP-3 chip, so that salient feature maps (e.g. spatial and temporal edges) can be generated, perhaps emulating the function of the retina. This filtered stimulus can then be passed on to the BG model. This could lead to a substantial part of the vertebrate oculomotor system model implemented on the SCAMP-3 system.

III. Experimental Results

The adapted BG model has been implemented on the SCAMP-3 processor array. A leaky integrator array is updated in 9 instruction-cycles, Gaussian connectivity is calculated in 2 instruction-cycles and the entire model takes 2996 instruction-cycles. Each instruction-cycle takes 800ns. Input is presented as a series of frames. For each new frame, the entire model is updated. To input data to the SCAMP-3 chip, a presentation is made of varying visual stimuli, shown on a monitor and viewed by a smart camera system [9] incorporating the SCAMP-3 device.

A series of inputs were presented to the system, representing different saliencies. Input consisted of patches of uniform luminance that varied in time, with a view to modelling four diffuse channel-sets with uniform input in each set. The SCAMP-3 BG implementation behaves in a

Table 1 - Comparison of SCAMP-3 and WORKSTATION model behaviour

Connectivity	Input Stimulus	SCAMP-3 Model Output (gain/offset adjusted)	WORKSTATION Model Output	Average Neuron Activity Difference
One-to-one				0.015
Gaussian				0.011

phenomenologically correct way. The regions of largest input are 'selected' (their inhibitory output activity is reduced from the quiescent state), whilst the remaining regions have increased output (inhibition is increased still further). Two versions of the BG model were implemented, one with Gaussian connectivity, and one with only one-to-one connectivity. Typically, the Gaussian model produced a less defined output, but behaved less erratically than the one-to-one model.

For a performance comparison, a computer model was constructed. This model was adapted to match the dimensions of the SCAMP-3 chip; however the weights and output functions were unaltered. The same scaling mechanism was also applied to the output of the STN, and the time-step was adapted so the models had similar response times. Effort was put into building the most optimised model possible. The workstation used featured a 1.8GHz Dual-Core AMD Opteron processor with 1GB of RAM, and 2MB of cache memory. The simulation was configured to run on a single core with real-time process scheduling, and make full use of the SSE architecture available. The simulation code was written in both C++ and assembly language. The input to the simulation came from a series of images recorded by the SCAMP-3 system, the same image input it used when executing the BG model. This is the only way the same input can be applied to both systems. Gaussian blurring was implemented as a 7x7 convolution matrix operation repeated 61 times.

The SCAMP-3 model and the WORKSTATION model behaved qualitatively in the same way, with the exception of linear scaling (gain and offset) present in the SCAMP-3 implementation due to the systematic errors of analogue processors. This gain and offset has been adjusted for the following analysis, using fixed values. Due to the noise and mismatch in analogue processors, the SCAMP-3 output is also somewhat more noisy, however both implementations produce very similar results. A snapshot of the model behaviour of both implementations is shown in Table 1. This table shows both input and output arrays (i.e. Input and GPi/SNr layers) for each platform. Note that the outputs represent inhibitory signals. The darker regions indicate lower inhibitory output, therefore "selecting" more salient (brighter) input. A highly salient input will increase the inhibition of competing channels. For a quantitative comparison the *average neuron activity difference* is a root-mean-square (rms) error between the SCAMP-3 and the WORKSTATION model output, calculated across all the neurons in the output layer over 1000 frames of input sequence. The average neuron activity difference is below 1.5% of the maximum signal value. This indicates that the SCAMP-3 architecture can satisfactorily match the arithmetic precision of the WORKSTATION in this application.

The relative effectiveness of the SCAMP-3 implementation, as compared to the workstation is evaluated through several metrics. It is interesting to look at the total time of computation; the time spent performing just neural

Table 2 - Performance Results

	Connectivity	Total Neural Computation Time (ms)	Total System Computation Time (ms)	Total Task Time (ms)
SCAMP-3	ONE-TO-ONE	2.399	0.022	2.410
SCAMP-3	GAUSSIAN	2.416	0.022	2.425
WORKSTATION	ONE-TO-ONE	1268.0	15732.0	17000.0
WORKSTATION	GAUSSIAN	67957.0	9043.0	77000.0

computation; and the time spent performing system computations (e.g. file I/O, screen updates, etc.). One thousand stimulus updates were presented to the models. It is assumed that every input layer neuron changes per stimulus. The BG model is updated after each stimulus change. Neural computation is a measure of how much time is spent updating neurons and distributing neuron output. System computation measures how much time is spent acquiring the stimulus, storing the results and maintaining model execution. These results are shown in Table 2.

The results show that the SCAMP-3 architecture is much more efficient at executing the basal ganglia model. This is particularly true for the Gaussian projection model, which performs at five orders of magnitude over the workstation counterpart.

The amount of system time required by both platforms indicates that the SCAMP-3 system is performing more efficiently. The workstation has to maintain memory and the order at which tasks are presented to the processor, resulting in significant overhead costs. The SCAMP-3 stores the data entirely within its processors, leaving only the transmission of frame data as a bottleneck. Typically, this bottleneck would be removed, as the SCAMP-3 system would most likely return parametric information, such as regions of most intensity.

The SCAMP-3 chip is a low power device, and when running constantly, consumes a maximum of 250mW. This implies that for the Gaussian model, one update takes about 600μJ to perform. The workstation's main processor consumes about 95W; meaning one update takes over 10kJ. This implies that the energy efficiency of the SCAMP-3 implementation is seven orders of magnitude that of the workstation. Of course, it needs to be acknowledged that the SCAMP-3 chip operates at lower accuracy than the WORKSTATION. However, the absolute numerical accuracy may be not that critical for many applications.

IV. Conclusions

We have demonstrated a large and complex leaky integrator neural network, associated with a biologically constrained model implemented on a single chip cellular processor array. The array enhances processing performance by up to five orders of magnitude, when compared to a high performance workstation. This speed-up results from the spatial arrangement and parallel nature of the processors which enabled traditional processing bottlenecks to be removed. This was particularly true for simulations which required Gaussian connectivity between neuron layers – a typical feature of biologically plausible models.

It has been shown that the SCAMP-3 can substantially out-perform a workstation at processing retinotopic leaky integrator neural networks, not just in terms of performance, but also in power consumption. This low-power neural processing would be ideal for embedded robotic applications and portable intelligent systems. The techniques presented here could be used to implement a variety of different networks.

Acknowledgements

This work has been supported by the EPSRC; grant number: EP/C516303.

References

[1] L. S. Smith, "Implementing neural models in silicon", Handbook of Nature-Inspired and Innovative Computing, Chapter 13, Part 5, A.Zomaya ed,Springer 2006

[2] T. Y. W. Choi, P. A. Merolla, J. V. Arthur, K. A. Boahen, B. E. Shi, "Neuromorphic Implementations of Orientation Hypercolumns", IEEE Transactions on Circuits and Systems, v52, n6, June 2005, 1049-59

[3] D. Balya, I. Petras, T. Roska, R. Carmona, A.R. Vazquez, "Implementing the multilayer retinal model on the complex-cell CNN-UM chip prototype", International Journal of Bifurcation and Chaos in Applied Sciences and Engineering, v14, n2, Feb. 2004, 427-51

[4] T. Watanabe, Y. Sugiyama, T. Kando, Y. Kitamura, "Neural network simulation on a massively parallel cellular array processor: AAP-2", IJCNN, 1989, 155-61 vol.2

[5] P.Dudek, "Implementation of SIMD Vision Chip with 128x128 Array of Analogue Processing Elements", IEEE International Symposium on Circuits and Systems, ISCAS 2005, Kobe, pp.5806-5809, May 2005

[6] G. E. Alexander, M. D. Crutcher, "Functional architecture of basal ganglia circuits: neural substrates of parallel processing", Trends in Neuroscience, 1990, v13, 266-272

[7] K. Gurney, T. J. Prescott, P. Redgrave, "A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour", Biol Cybern, 2001, v84, n6, 411-423

[8] J. M. Chambers, "Deciding where to look: A study of action selection in the oculomotor system", Ph.D. dissertation, University of Sheffield, 2007

[9] D.R.W.Barr, S.J.Carey, A.Lopich and P.Dudek, "A Control System for a Cellular Processor Array", IEEE International Workshop on Cellular Neural Networks and their Applications, CNNA 2006, pp.176-181, Istanbul, August 2006

[10] P.Dudek and P.J.Hicks, "A General-Purpose Processor-per-Pixel Analog SIMD Vision Chip", IEEE Transactions on Circuits and Systems - I, vol. 52, no. 1, pp. 13-20, January 2005