Brain Inspired Cognitive Systems
August 29 – September 1, 2004
University of Stirling, Scotland, UK

A NEUROMORPHIC SELECTIVE ATTENTION ARCHITECTURE WITH DYNAMIC SYNAPSES AND INTEGRATE-AND-FIRE NEURONS

Chiara Bartolozzi
Institute for neuroinformatics
UNI-ETH Zurich
Wintherthurerstr. 190, 8057, Switzerland
Email: chiara@ini.phys.ethz.ch

Giacomo Indiveri
Institute for neuroinformatics
UNI-ETH Zurich
Wintherthurerstr. 190, 8057, Switzerland
Email: chiara@ini.phys.ethz.ch

ABSTRACT

Selective attention is a process widely used by biological sensory systems to overcome the problem of limited parallel processing capacity: salient subregions of the input stimuli are serially processed, while non–salient regions are suppressed. We present an analog Very Large Scale Integration implementation of a building block for a multi–chip neuromorphic hardware model of selective attention. We describe the chip's architecture underlining the similarity between the circuits and biological neurons and synapses. We present experimental results showing the system's behavior as a function of its bias settings.

INTRODUCTION

Selective attention is one of the most powerful strategies used by biological systems, from which robotics and in general all artificial computation can take advantage. In a biological sensory system, selective attention acts as a dynamical filter that selects the most salient regions of the input, sequentially allocating computational resources, for analyzing the target. This strategy limits the computational demand respect to parallel processing. The selection of one between possible targets depends on its *'saliency'*; the saliency of a stimulus depends on its physical and semantic characteristics and on the relevance it has for the ongoing activity of the subject. There are two main pathways that determine the emergence of one 'winning' stimulus in the competition for saliency: one is stimulus-driven, bottom-up and task-independent, the other is goal-dependent, and acts in a slower top-down manner.

Much of the research focused on modeling the bottom-up aspect of selective attention, gave rise to software [1–4] and hardware models [5–8] based on the concept of *saliency map* [9]. Software models based on this concept account for many psychophysical and neurophysiological observations [10] and have features that could be used in practical applications. Hardware implementations of selective attention systems have the additional advantage of real time computation and compactness: they can be used for building artificial systems that interact with real world stimuli in real time, and can therefore be a powerful tool for studying computational properties of different types of selective attention models.

The concrete physical realization of these models has to take into account issues such as noise, limited resources and power availability, as well as fault tolerance, and robustness to variations in the input, very much like the brain has to. This will hopefully lead to a better understanding of the physical and computational mechanisms used by the brain to solve these problems, including details that might be overlooked in abstract models or computer simulations.

Here we present a VLSI device, the Selective Attention Chip (SAC), that can be used as a building block for hardware multi–chip sensory systems, based on selective attention models. Specifically the SAC represent a hardware implementation of a saliency-based computational model of the bottom-up dynamical form of selective attention [11]. The SAC was realized with Very Large Scale of Integration (VLSI) technology using neuromorphic circuits that directly map biophysical neuronal properties onto silicon. It employs a spike-based representation both

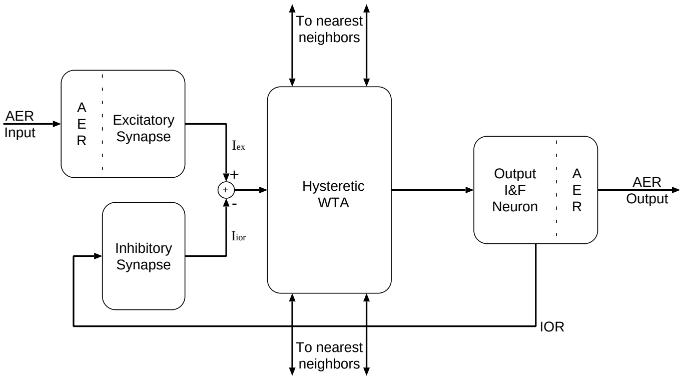

Figure 1. Block diagram of a basic cell of the 32 × 32 selective attention architecture.

for receiving input signals and for transmitting output signals to further processing stages. Its input signals are expected to arrive from a saliency map, topographically encoding local conspicuousness over the entire visual scene. Its output signals can be used in real time to drive motors of active vision systems or to select subregions of images captured from wide field-of-view cameras.

This chip is the evolution of a similar device previously proposed in [12]. This new device augments the previous one by implementing a larger array (32×32 cells as opposed to 8×8), by using a novel low-power spiking neuron circuit [13], and by using more advanced synaptic circuits with realistic dynamics and adaptation properties [14, 15].

In the next sections we describe the chip's architecture and its main circuital elements and show both behavioral simulation results and some preliminary experimental results that illustrate the effects of the new types of synapses and neurons on the selective attention dynamics.

THE SAC ARCHITECTURE

The SAC was fabricated using a standard $0.35\mu m$ CMOS technology; it contains an array of 32×32 cells laid out on a square grid; a single cell is $50.65 \times 32\mu m^2$ while the whole array occupies an area of $23447\mu m^2$. Each cell in the bidimensional array comprises an input circuit that models the dynamics of a biological excitatory synapse, generating Excitatory Post–Synaptic Currents (EPSCs), a hysteretic Winner–Take–All (WTA) competitive element [16], an output Integrate and Fire (I&F) neuron [13] and a feedback inhibitory synapse (see Fig. 1).

Input and output signals of the SAC are asynchronous digital pulses (spikes) that use an *Address Event Representation* (AER) [17]. The AER is inspired by cortical communication: it is based on asynchronous events (spikes) that encode the address of the sending neuron and carry the analog information in their temporal structure. This protocol allows multiple AER chips to communicate using spikes, just like the cortex, and can be used in multi–chip systems, with multiple senders and multiple receivers [18, 19]. Using this representation the SAC can exchange data, while processing signals in parallel, in real time [12]. The communication protocol used and the SAC's bidimensional architecture make it particularly suitable for processing visual inputs coming from artificial spiking retinas or cochleas.

Input spikes arriving for example from a silicon retina [20] or from a software based vision system [8] are integrated by the excitatory synapses of the array into excitatory analog current (see I_{ex} of Fig.1); the effect of a single spike on the integrated current depends on the synaptic weight V_w of Fig. 2(a). The initial weight of the synapse is set by an external voltage reference (V_{w0} of Fig. 2(a)), then as the synapse receives spikes (voltage pulses *pre*) the effective synaptic weight V_w decreases, in a way to model local gain control, reproducing *short time depression* dynamics observed in physiological recordings [21].

The integrated excitatory current is sourced into the correspondent WTA cell that competes with the other cells by means of lateral excitatory and inhibitory connections. The spatial extent of the competition can be set by the strength of these lateral connections; in particular we can set global competition, allowing only one cell to win, or we can have local competition, with multiple spatially distant winners [16].

As soon as a WTA cell wins the competition it sources a fixed amount of current into the membrane capacitance of the adaptive low power I&F neuron. The spiking frequency of the I&F neuron is monotonic with its input current. The adaptation neuron's mechanism decreases the neuron's firing rate with time [13].

The output spikes go to an arbitration circuit that sends the address of the winning pixel to the AER bus and, in parallel, to the corresponding inhibitory synapse that is responsible for generating the inhibitory current I_{ior} (see Fig.1); this current is subtracted from the input excitatory current I_{ex}, therefore the net input current to the winning cell decreases until a different cell is eventually selected as winner. This negative feedback mechanism is known as Inhibition of Return (IOR), it allows the network to deselect the winning cell and switch between inputs with different salience.

The SAC has been designed with tunable parameters that allow to modify the strength of synaptic contributions, the dynamics of synaptic short term depression and of neuronal adaptation, as well as the spatial extent of competition and the dynamics of IOR. All these parameters enrich the dynamics of the network that can be exploited to model the complex selective attention scan path.

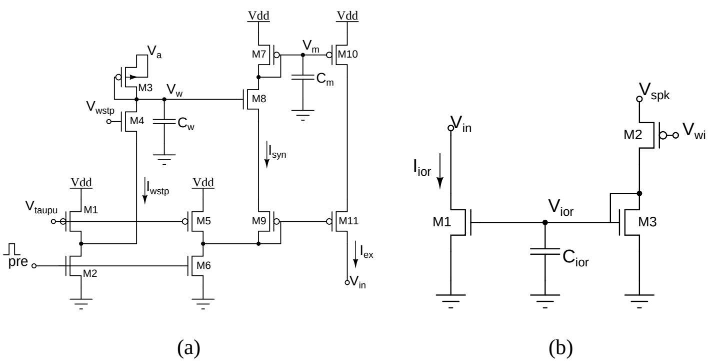

Figure 2. (a) Excitatory synapse circuit. Input spikes are applied to M1, and transistor M11 outputs the integrated excitatory current I_{ex}. (b) Inhibitory synapse circuit. Spikes from the local output neurons are integrated into an inhibitory current I_{inh}.

BEHAVIORAL SIMULATIONS AND EXPERIMENTAL DATA

In order to assess the dynamical properties added by synaptic short term depression and spiking frequency adaptation, we simulated the behaviour of 2 pixels, using the analytical equations derived from the circuits. In addition we performed some preliminary experiments to characterize the chip implemented.

Excitatory Synapse

The current mirror integrator circuit [22] in the excitatory synapse integrates the incoming spikes, decreasing the gate voltage V_m of the output transistor. We can derive the time course of V_m during the spike from Kirchoff's current law and from the transistor's weak inversion equations:

$$I_{syn} = -C_m \frac{dV_m}{dt} + I_{0_p} e^{\frac{\kappa(V_{dd}-V_m)}{U_T}} \tag{1}$$

Where I_{0_p} is the transistor's dark current, U_T is the thermal voltage, κ is the transistor subthreshold slope factor and V_{dd} is the power supply. Integrating Eq. 1:

$$V_m(t) = \frac{U_T}{\kappa} \ln\left((e^{\frac{\kappa V_{m0}}{U_T}} - \frac{I_{0_p}}{I_{syn}} e^{\frac{\kappa V_{dd}}{U_T}}) e^{-\frac{\kappa I_{syn}}{U_T C_m} t} + \frac{I_{0_p}}{I_{syn}} e^{\frac{\kappa V_{dd}}{U_T}} \right) \tag{2}$$

During a spike the voltage V_m is decreased by an amount determined by I_{syn} that depends exponentially on the synaptic weight V_w. The short term depressing part of the synapse (transistors $M1 - -M4$) of Fig. 2(a) decreases V_w with each spike. To a first order approximation during a spike the synaptic weight decreases linearly:

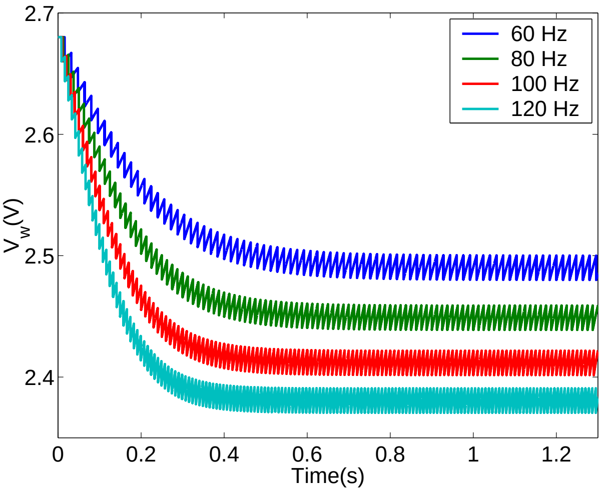

Figure 3. Short term depression simulations. The synaptic weight Vw is plotted for input spike trains of different frequency. The higher the frequency, the lower is the steady-state depressed value.

$$V_w(t) = V_{w0} - \frac{I_{wstd}}{C_w} t \tag{3}$$

During the time interval between spikes transistors $M3$ and $M7$ in the synapse of Fig. 2 tend to restore V_w and V_m respectively. In this case the synapse has no input and V_m can be obtained integrating Eq. 1, for $I_{syn} = 0$:

$$V_m(t) = \frac{U_T}{\kappa} \ln(e^{\frac{\kappa V_{m0}}{U_T}} + \frac{\kappa I_{0_p}}{U_T C_m} e^{\frac{\kappa V_{dd}}{U_T}} t) \tag{4}$$

In the same way V_w is obtained integrating $\frac{dV_w}{dt} = \frac{I_{M1}}{C_w}$:

$$V_w(t) = \frac{U_T}{\kappa} \ln \frac{\kappa I_{0_p}}{U_T C_w} e^{\frac{\kappa V_a}{U_T}} t + e^{\frac{\kappa V_{w0}}{U_T}} \tag{5}$$

In Fig. 3 we show the variation of the synaptic weight V_w when the synapse is stimulated with constant spike trains for increasing input firing rates, the steady state of depression decreases with spiking frequency of the input. In Fig. 4 we show the variation of the synaptic weight V_w when the synapse is stimulated with a constant spike train for two values of the depressing bias V_{wstd}, both the steady state and the time course of the synaptic weight can be changed.

WTA

The hysteretic WTA cell compares its input current to the current of the winning cell plus an hysteretic current I_{hyst}; the

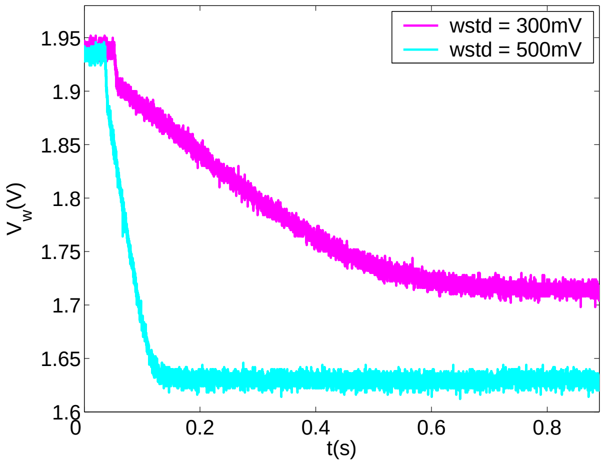

Figure 4. Short term depression of the excitatory synapse. Data obtained from the chip showing V_w for different values of short term depression settings (V_{wstd}). The higher V_{wstd}, the lower is the steady state value of the synaptic weight

hysteretic current gives to the currently winning cell a competitive advantage, implementing a sort of short time memory that could be useful for tracking salient patterns. The input current to the WTA cell is the sum of the positive current sourced by the excitatory input synapse I_{ex}, and the negative current subtracted by the IOR inhibitory synapse I_{ior}.

In Fig. 5 we show the effect of alternatively stimulating two pixels: the activity of the two output neurons alternates indicating which pixel is winning the competition for saliency.

I&F neuron

The I&F neuron integrates its input current until the integrated membrane voltage crosses a threshold. At that point the neuron generates a spike and the membrane voltage is reset. We can model the subthreshold time course of V_{mem} by:

$$C_{mem}\frac{d}{dt}V_{mem} = I_{wta} - I_{leak} + I_{fb} - I_{adap} \tag{6}$$

where the net current in input is given by the current sourced by the WTA cell I_{wta}, minus a leakage current

$$I_{leak} = I_{0_n} e^{\frac{\kappa}{U_T}V_{lk}}(1-e^{-\frac{V_{mem}}{U_T}}) \tag{7}$$

a feedback current

$$I_{fb} = I_1 e^{-\kappa^2\frac{V_{sf}}{U_T}} e^{\kappa^2\frac{V_{mem}}{U_T}} \tag{8}$$

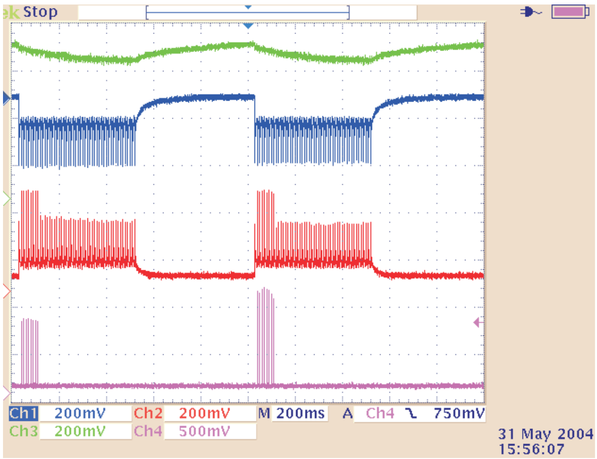

Figure 5. Experimental data for the WTA circuit:pixel $(0,0)$ and pixel $(1,1)$ are stimulated with a constant spike train at $100Hz$ that switches between the two each second. We show the output screen of the digital oscilloscope for pixel $(0,0)$: the top trace represents the weight of the short-term depressing synapse, the second trace is the voltage in the current mirror integrator of the excitatory synapse (V_m of Fig.3), the third trace is logarithmically proportional to the total current in input to the WTA cell, and the bottom trace is the neuron's membrane voltage. Every time the pixel $(0,0)$ is stimulated the synapse integrates the incoming spikes, the weight adapts, the current in the WTA cell increases and the $I\&F$ neuron spikes. When the synaptic weight is sufficiently depressed the neuron stops firing and the other neuron wins the competition for salience.

and an adaptive current that increases for each spike,

$$I_{adap} = I_0 e^{\kappa\frac{V_{a0}}{U_T}} e^{\kappa\gamma\frac{V_{mem}}{U_T}}(1-e^{-\frac{V_{mem}}{U_T}}) \tag{9}$$

thanks to I_{adap} the effect of a constant current applied to the neuron decreases with time, resulting in a decrease of the output firing rate that will affect the dynamics of the IOR mechanism.

Results

Even the when stimulating only two pixels the network shows an interesting dynamic behaviour enriched mostly by the introduction of the short term depression in the excitatory synapse. In the behavioural simulations we stimulated two pixels with constant spike trains of different frequencies; the change in the synaptic weight depends on the input frequency as shown in Fig. 3. This effect is a useful feature that equalizes the inputs coming from noisy sources as spiking retinas [20]. This behaviour enhances responses to stimuli that change in time rather than to constant or slow stimuli. In Figure 6 and 7 we show the change in the synaptic weight and in the synaptic output current respectively for a 4 seconds simulation, where the frequency of the input spike trains changes every second: Pixel one is stim-

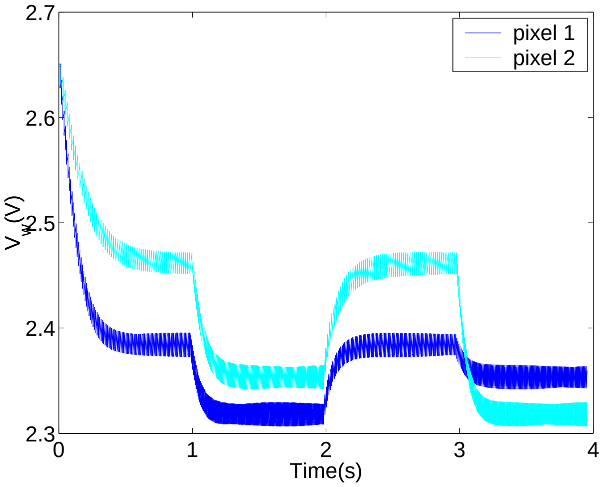

Figure 6. Synaptic weight of the two pixels. Pixel one is stimulated at $100Hz$, then at $150Hz$, $100Hz$, $120Hz$, pixel two receives $60Hz$, $120Hz$, $60Hz$ and eventually $150Hz$. V_w depends on the absolute value of the frequency.

ulated at $100Hz$, then at $150Hz$, $100Hz$ and $120Hz$, while pixel two receives $60Hz$, $120Hz$, $60Hz$ and eventually $150Hz$. The value of the synaptic weight depends on the absolute value of the frequency. The current peaks in correspondence to relative changes in the input: the synapse can be seen as an high pass filter, since it enhances the input's temporal variations. In this experiment the network shows an even behaviour: after a transitory, due to the stimulus variation, it starts to switch between the most salient and the second most (less in our case) salient input, thanks to a balance between IOR and hysteresis.

A similar experiment has been performed on the chip and the result is shown in Fig. 5, in this case the IOR was switched off and the dynamics of the WTA depends only on the short term depression of the synapse.

CONCLUSIONS

In this paper we presented a neuromorphic device implementing a Winner–Take–All network. This device is designed to be a part of a multi–chip system for Selective Attention: via AER communication system it can be interfaced to silicon spiking retinas and to software implementations of associative memories.

We have shown that the new synaptic circuits can equalize the input to the competitive network, therefore it can cope with noisy inputs.

The prohibitive CPU simulation times for larger networks simulations didn't allow us to explore the possible additional features introduced by short term depression and spike frequency adaptation. The real time measurements allowed by the physical realization of the chip are certainly a more powerful method to explore the network behaviour by changing its parameters. The preliminary experiments confirmed the simulation's results and will be extended with the introduction of IOR, adaptation and lateral coupling among the nearby cells.

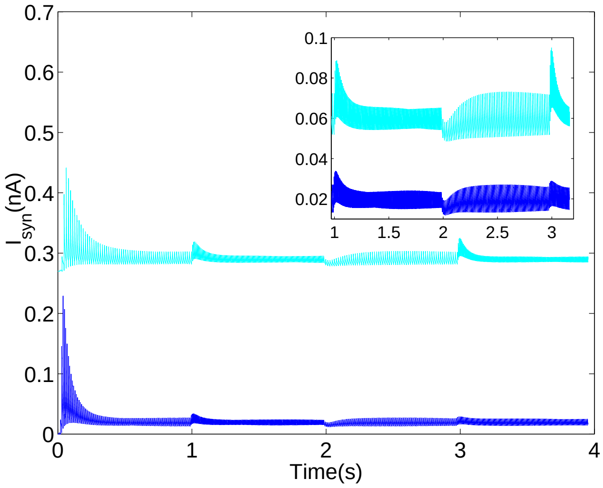

Figure 7. Synaptic output current, for the same simulation as in Fig. 6. The current shows peaks in correspondence of the stimulus variations, the amplitude of the peak is related to the value of the input frequency step.

ACKNOWLEDGMENT

This work was inspired by interactions with the participants of the Neuromorphic Engineering Workshop (http://www.ini.unizh.ch/telluride) and was supported by the EU grant ALAVLSI (IST–2001–38099) and the Italian CNR (Centro Nazionale delle Ricerche) fellowship 203.22. I further wish to thank Matthias Oster for the implementation of the DAC board software.

REFERENCES

[1] Itti, L., Niebur, E., and Koch, C., 1998. "A model of saliency-based visual attention for rapid scene analysis". IEEE Trans. on Pattern Analysis and Machine Intelligence, **20** (11) , pp. 1254–1259.

[2] Bosch, H., Milanese, R., and Labbi, A., 1998. "Object segmentation by attention-induced oscillations". In Proc. IEEE Int. Joint Conf. Neural Networks, vol. 2, pp. 1167–1171.

[3] Trahanias, P., Velissaris, S., and Garavelos, T., 1997. "Visual landmark extraction and recognition for autonomous robot navigation". In Proc. IEEE Int. Conf. Intelligent Robots and Systems IROS '97, vol. 2, pp. 1036–1043.

[4] Baluja, S., and Pomerleau, D., 1997. "Expectation-based selective attention for the visual monitoring and control of a robot vehicle". *Robotics and Autonomous Systems Journal*, **22** , pp. 329–344.

[5] Brajovic, V., and Kanade, T., 1998. “Computational sensor for visual tracking with attention”. IEEE Journal of Solid State Circuits, **33** (8) Aug. , pp. 1199–1207.

[6] Horiuchi, T., and Koch, C., 1999. “Analog VLSI-based modeling of the primate oculomotor system”. *Neural Computation,* **11** , pp. 243–265.

[7] Morris, T. G., Horiuchi, T. K., and DeWeerth, S. P., 1998. “Object-based selection within an analog VLSI visual attention system”. IEEE Trans. on Circuits and Systems II, **45** (12) , pp. 1564–1572.

[8] Indiveri, G., 2000. “Modeling selective attention using a neuromorphic analog VLSI device”. Neural Computation, **12** (12) December , pp. 2857–2880.

[9] Koch, C., and Ullman, S., 1985. “Shifts in selective visual-attention – towards the underlying neural circuitry”. Human Neurobiology, **4** (4) , pp. 219–227.

[10] Itti, L., and Koch, C., 2001. “Computational modeling of visual attention”. *Nature Neuroscience Review,* **2** , pp. 194–204.

[11] Niebur, E., and Koch, C., 1998. “Computational architectures for attention”. In *The Attentive Brain*, R. Parasuraman, Ed. MIT Press, pp. 163–186.

[12] Indiveri, G., 2001. “A neuromorphic VLSI device for implementing 2-D selective attention systems”. IEEE Trans. on Neural Networks, **12** (6) November , pp. 1455–1463.

[13] Indiveri, G., 2003. “A low-power adaptive integrate-and-fire neuron circuit”. In Proc. IEEE International Symposium on Circuits and Systems, IEEE.

[14] Rasche, C., and Hahnloser, R., 2001. “Silicon synaptic depression”. Biological Cybernetics, **84** (1) , pp. 57–62.

[15] Boegerhausen, M., Suter, P., and Liu, S.-C., 2003. “Modeling short-term synaptic depression in silicon”. Neural Computation (15) Feb , pp. 331–348.

[16] Indiveri, G., 2001. “A current-mode hysteretic winner-take-all network, with excitatory and inhibitory coupling”. Analog Integrated Circuits and Signal Processing, **28** (3) September , pp. 279–291.

[17] Mahowald, M., 1994. *An Analog VLSI System for Stereoscopic Vision*. Kluwer, Boston.

[18] Dante, V., and Del Giudice, P., 2001. “The PCI-AER interface board”. In 2001 Telluride Workshop on Neuromorphic Engineering Report, A. Cohen, R. Douglas, T. Horiuchi, G. Indiveri, C. Koch, T. Sejnowski, and S. Shamma, Eds., pp. 99–103. http://www.ini.unizh.ch/telluride/previous/report01.pdf.

[19] Deiss, S. R., Douglas, R. J., and Whatley, A. M., 1998. “A pulse-coded communications infrastructure for neuromorphic systems”. In *Pulsed Neural Networks*, W. Maass and C. M. Bishop, Eds. MIT Press, ch. 6, pp. 157–178.

[20] Kramer, J., 2002. “An ON/OFF transient imager with event-driven, asynchronous readout”. In Proc. IEEE International Symposium on Circuits and Systems.

[21] Abbott, L., Sen, K., Varela, J., and Nelson, S., 1997. “Synaptic depression and cortical gain control”. Science, **275** (5297) , pp. 220–223.

[22] Boahen, K., 1998. “Communicating neuronal ensembles between neuromorphic chips”. In *Neuromorphic Systems Engineering*, T. S. Lande, Ed. Kluwer Academic, Norwell, MA, pp. 229–259.