Diss. ETH No. 17156

Selective attention in silicon: from the design of an analog VLSI synapse to the implementation of a multi–chip system

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
(ETH ZURICH)

for the degree of
Doctor of Sciences

presented by
CHIARA BARTOLOZZI
Dipl.–Eng. Università di Genova
born September 12, 1977
citizen of Genova, Italy

accepted on the recommendation of
Prof. Dr. Rodney J. Douglas, examiner
Dr. Giacomo Indiveri, co-examiner
Prof. Dr. Peter König, co-examiner

2007

Abstract

The basic elements of the cortical neural substrate and of current silicon technology obey similar physical principles; the implementation of systems based on each of these two ”technologies” also faces similar constraints. The foundation of neuromorphic engineering is to recognize and exploit such similarities, and map the properties of neural computation on to silicon to implement new types of computing devices. This approach leads to the understanding of some of the principles that shape neural computation, and to the implementation of efficient devices which can interact with the real world in real time. The neuromorphic strategy has particular relevance for applications where biological systems outperform classical digital computers, such as the task of perception, where the system must process noisy and ambiguous stimuli to produce appropriate behavioral responses. The efficient and compact devices developed through this approach are especially suited for integration into autonomous artificial systems; an example is the design of ”smart” sensors for robotic platforms. Progresses in this area could eventually lead to the realization of new prosthetic devices, that, interfacing and processing information in a way similar to biological nervous systems, could be naturally interfaced to biological nervous systems.

The work presented in this thesis ranges from ”morphing” properties of synaptic transmission on to silicon, to the realization of a Selective–Attention Chip (SAC), integrated in a multi–chip system which implements a model of visual selective attention, capable of operating in the real world, in real time.

The silicon synapse described in this thesis produces realistic postsynaptic currents in response to presynaptic spikes; it offers the possibility of emulating the functionality of NMDA and conductance–based synapses in the same framework; most importantly, it is compatible with circuits that model plasticity on short and long time scales, ranging from fast adaptation, to spike–based learning, to homeostatic plasticity. Each of the properties emulated by the proposed synapse circuit has a specific role in neural computation, and therefore enriches the vocabulary of computational primitives that now can be included and studied in neuromorphic systems.

The chip developed in this thesis makes extensive use of this synaptic circuit, and sequentially selects the most salient regions of the input stimuli. The selection process, known as *selective attention*, is the strategy used by natural perceptive systems to cope with the enormous amount of information received, in face of their limited processing capabilities. Artificial sensory systems can also benefit from such a strategy.

Analogous to its biological counterpart, the selection mechanism implemented by the SAC emulates a typical emergent computational property of recurrent neural networks, that arises from cooperation and competition between computational units, which extracts information from the relative activity of each unit. The SAC functionality is founded on the dynamical interaction of the various components of the selective attention process, for which the synapse circuit developed in this work is a crucial element.

The multi–chip system proposed in this thesis comprises the SAC and a neuromorphic

transient imager, mounted on an controllable platform. The imager responds to local variations of the input luminance over time. In such a visual system, the attentional selection is based on the temporal variation of the stimulus contrast. The activity of the SAC is used to orient the imager toward the selected stimulus. Such a system expresses an elaborate behavior, with a mixture of tracking of selected targets and attentional shifts. The results obtained demonstrate the usefulness and potential application of neuromorphic sensors and post–processing devices such as the SAC to artificial perceptive systems, in the context of robotic systems.

The physical realization of the perceptive system proposed in this thesis is also relevant in the context of selective attention research. Specifically it can be used as a tool for validating hypotheses arising from experimental observations of biological systems and computational models.

Prefazione

Gli elementi costitutivi del substrato neurale e dell'attuale tecnologia in silicio obbediscono a principi fisici analoghi; anche la realizzazione concreta di sistemi basati su entrambe le "tecnologie" deve far fronte a limitazioni e problematiche simili. Il fondamento della ricerca in ingegneria "neuromorfa" si basa sul riconoscere e sfruttare tali similitudini, per trasportare su silicio le proprietà della computazione neurale e realizzare nuovi tipi di dispositivi computazionali. Questo approccio consente di rivelare alcuni dei principi che hanno contribuito a modellare la computazione neurale e di costruire dispositivi efficienti, capaci di interagire con il mondo reale, in tempo reale.

La strategia neuromorfa ha particolare rilevanza nelle applicazioni in cui i sistemi biologici risultano più efficienti dei "classici" computer basati sulla computazione digitale, come per esempio in compiti che coinvolgono la percezione, ossia in applicazioni in cui i sistemi devono processare stimoli ambigui e rumorosi e produrre comportamenti appropriati. Uno sbocco naturale di questo approccio è lo sviluppo di dispositivi elettronici compatti ed efficienti, adatti ad essere incorporati in sistemi artificiali autonomi, come per esempio sensori "intelligenti" per piattaforme robotiche. Il progredire di questa linea di ricerca porterà infine alla creazione di nuovi tipi di protesi che, interagendo con il mondo reale in modo simile ai sistemi nervosi biologici, potranno interfacciarsi con essi in modo naturale.

Il lavoro presentato in questa tesi parte dalla trasposizione su silicio delle proprietà della trasmissione sinaptica per arrivare alla costruzione del "Selective–Attention" Chip (SAC), integrato in un sistema multi–chip che realizza un modello visivo di attenzione selettiva, capace di interagire con il mondo reale, in tempo reale.

Il circuito sinaptico presentato riproduce le correnti caratteristiche che attraversano la membrana di un neurone post–sinaptico all'occorrenza di uno o più potenziali d'azione pre–sinaptici. Tale circuito può essere corredato di circuiti addizionali, che ne estendono la funzionalità, emulando le caratteristiche delle sinapsi NMDA e la dipendenza delle correnti post–sinaptiche dal potenziale di membrana. Fondamentale è poi la possibilità di includere nel modello circuiti che attuano la dinamica del peso sinaptico a breve e a lungo termine, spaziando dall'adattamento, all'apprendimento, fino alla plasticità omeostatica.

Ciascuna delle proprietà emulate dal nuovo circuito sinaptico proposto ha un suo ruolo specifico nella computazione neurale, e quindi arricchisce il vocabolario delle primitive computazionali che possono essere incluse e studiate nei sistemi neuromorfi.

Il chip sviluppato in questa tesi include il circuito sinaptico, che ne è un elemento essenziale, e seleziona in modo sequenziale le regioni più salienti dello stimolo in ingresso. Tale funzione, nota come "Attenzione Selettiva", costituisce la strategia utilizzata dai sistemi percettivi biologici, per gestire l'enorme quantità di dati sensoriali, rispetto ai limiti della capacità di computazione parallela. Da questo tipo di strategia possono trarre vantaggio anche i sistemi percettivi artificiali.

Il processo di selezione realizzato dal SAC si basa sulla cooperazione e competizione tra unità computazionali, che estraggono l'informazione dalla relazione tra l'attività di ogni

unità, modellando una proprietà emergente delle reti neurali con connettività ricorrente. Per la funzionalità del SAC è cruciale l'interazione dinamica dei vari elementi, in cui il circuito sinaptico sviluppato in questo contesto è un elemento chiave.

Il sistema finale proposto in questa tesi comprende il SAC ed una retina neuromorfa, montata su un attuatore, che risponde alle variazioni locali di intensità luminosa nel tempo. In tale sistema visivo, la selezione attentiva è determinata dalla variazione temporale del contrasto degli stimoli visivi. L'attività del SAC viene poi usata per orientare la retina verso lo stimolo selezionato. Questo sistema genera un comportamento complesso, con un alternarsi dinamico di tracking di stimoli selezionati e selezione di nuovi stimoli.

I risultati ottenuti dimostrano l'utilità e le potenzialità dell'uso di sistemi percettivi basati su sensori neuromorfi e chip come il SAC, nel contesto della realizzazione di sistemi robotici che richiedono l'interazione con il mondo reale, in tempo reale.

La realizzazione fisica di un sistema percettivo come quello proposto in questa tesi ha rilevanza anche nel contesto della ricerca sull'attenzione selettiva; in particolare può essere usato come strumento per validare ipotesi che nascono sia da esperimenti su sistemi attentivi biologici, sia da modelli computazionali.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and overview

1.1 The neuromorphic quest

Any brain, from the human brain to the smallest insect's brain (Srinivasan et al., 1996), outperforms computers on tasks involving interaction with the real world, where ambiguous inputs have to be interpreted to initiate actions. Guyonneau et al. (2006) have shown that human subjects can judge if an animal is present in a natural scene picture and make an eye movement toward the animal in less than 250ms; this task is performed as quickly when the images are rotated, showing a striking example of the degree of rotation, size, and point of view invariance of the human visual system. This simple task for humans cannot be achieved by computer vision, which is still struggling with problems such as object–background segregation and rotation invariance, among others.

In terms of energy consumption, Sarpeshkar (1998) estimates the efficiency of brain computation as about $3{\cdot}10^{14}$ operations per joule. Conversely, a modern microprocessor performs at about $6.25{\cdot}10^{6}$ operations per joule[1]. The supercomputer BlueGene/L, designed by IBM (Gara et al., 2005) and awarded as top supercomputer in 2006[2], can reach a theoretical peak of $367{\cdot}10^{15}$ floating point operations per second with a power consumption of about 10^{6}W — this corresponds to a performance of 10^{9} operations per joule, still several orders of magnitude lower than the brain efficiency. BlueGene's performance is achieved by parallelizing 131072 processors; — Fig. 1.1 gives an idea of the dimensions of this computing system, which needs rooms and special cooling devices.

Indeed a computer comprising very fast and precise elements fails in solving tasks that appear to be trivial for even the smallest brain, which is built with slow, imprecise and inhomogeneous elements.

The reasons for this discrepancy are not yet fully understood. One of the quests of neuromorphic engineering is understanding these reasons, and implementing more efficient devices by exploiting the strategies developed by the brain and the physics of a silicon substrate.

The most obvious difference between computers and brains is that the first is based on Boolean logic, and computes using an alphabet that comprises only "0" and "1", largely in a serial fashion. Conversely, brain computation is highly parallel and is performed by neurons and synapses, whose primitives are based on analog variables.

Mead (1990), the father of neuromorphic engineering, attributed the reasons of the out–performance of the brain over digital computers to the use of elementary physical phenom-

[1] These data refer to 1998, they are mentioned here to give a rough idea of the scale of the difference. In the past years technology has improved, but has not yet reached the scale of integration, range of power consumption or the capabilities of brains to analyze and react to real world stimuli.

[2] http://www.top500.org/

Figure 1.1: Blue Gene/L is housed at the Terascale Simulation Facility (TSF) at Lawrence Livermore National Laboratory: Two supercomputers occupy an entire floor of the 23,504m^2 building (Blue Gene/L machine floor space is 232m^2), a second floor is devoted to the housing of cooling devices: "A total of 28 air–handling units blow cool air up to the second level, each at a rate of 80,000 cubic feet per minute". Adapted from *http://www.llnl.gov/asc/asc_index.html*

ena as computational primitives; VLSI technology — provided that it is used in the weak inversion, or subthreshold, domain — shares the same primitives of neural computation, that are not exploited by digital technology.

In both CMOS transistor physics and in nervous tissue, the state variables are analog signals. In the nervous system, such variables are represented by ion concentrations that translate into electro–chemical potentials. In transistors, the charges are carried by electrons and holes. Both electronics and the nervous system base their functionality on the existence of energy barriers. In biology, the barrier is constituted by selectively permeable cell membrane. In electronics, it is built by the difference in the band gap between silicon and silicon dioxide. Accumulation of charge carriers in thermal equilibrium across an energy barrier results in a Boltzmann distribution of their energies (Mead, 1990; Grattarola and Massobrio, 1998), that in turns results in an exponential dependence of the current across the barrier on the voltage difference applied.

In both electronics and neural tissue, information is stored by accumulation of charge, and computation is naturally performed with currents, which can be added in space (Kirchhoff current law) and integrated in time by capacitances; non linearities such as saturation and thresholding are intrinsic in both media.

Biological systems and VLSI share not only the same computational primitives, but also similar constraints. Both have a finite power supply and need to optimize energy consumption. In VLSI it is critical when building portable devices to increase battery life and reduce the need for cooling; the same is true for living creatures, which need to perform with limited resources and consequently developed energy–efficient designs and computational codes (Laughlin and Sejnowski, 2003).

Equally important is the need for space optimization, for both costs relative to the use of a physical medium and "portability". A critical issue influencing the integration level of both systems is wiring. The brain has optimized wiring (Laughlin and Sejnowski, 2003;

Mead, 1990) by exploiting local computation, and the co–localization of state variables with computational structures. This strategy is effective both for saving space and for decreasing power consumption. In digital computers, where the state variables are stored in different physical locations from the site of computation, the majority of energy is spent in charging the wires transferring information, not in the gates of the computing transistors. A factor of 100 in efficiency over classical digital computing can be gained by implementing the same strategy as the neural substrate (Mead, 1990), in which computation and memory storage are co–localized.

Another factor constraining the development of computation in systems interacting with the real world is the dynamic range of input signals, whose amplitude can vary up to 10 orders of magnitude. The elegant solutions to this problem — gain control, adaptation and relative rather than absolute computation — are present at all levels of brain computation, from the photoreceptors in the retina to the neurons in cortex (Ohzawa et al., 1985; Kandel et al., 2000). Successful replications of these properties have been implemented in VLSI, the most brilliant being the adaptive photoreceptor designed by Delbrück and Mead (1995). This led to the implementation of a silicon retina (Lichtsteiner et al., 2006a) working in a wide range of luminance conditions, from moon–light to full sun.

Probably the most striking common constraint that both biology and analog subthreshold silicon face is the intrinsic inhomogeneity and low precision of their fundamental constituents. Douglas et al. (1995) attribute the main difference between digital and neural computation, and hence the reason for superiority of the latter in real world interaction, to the fundamental difference in strategy for obtaining precision in computation. Digital computing relies on more homogeneous components and deals with signal restoration at the lowest level, by fully restoring each bit at every step of the computation. In neural computing the analog values cannot be attracted to the closest "permitted" value, rather they are restored in a collective fashion by recurrent connectivity, which relates the signals to the context of the activity of many surrounding computational units. Precision is obtained at the collective, rather than individual level. Connectivity between units is the foundation for cooperation and competition between related signals, and the same mechanisms that restore the signals may lead to the extraction of useful information from noisy and ambiguous signals (Douglas et al., 1994).

Biology copes with intrinsic inhomogeneities of its substrate, failures of its components or changes in the environment, by dynamically changing the properties of single computational units (Turrigiano, 1999). This property makes neural computation robust and fault tolerant, and naturally leads to systems that learn and adapt to their environment (Mead, 1990).

The computational benefit of this approach is clear when comparing the overall robustness of digital computers against that of brains. A system with high enforced precision, both at the component level and at each stage of computation, is less robust and less tolerant to component degradation or failure than an inhomogeneous system with intrinsic low precision. The first approach separates computation from signal restoration, with very high resource cost; it is effective for high precision numeric computation, but the failure of one component causes the failure of the whole system (Sarpeshkar, 1998); additionally it is ineffective in solving ill–posed problems, where relevant information must be extracted from noisy and ambiguous signals. In such a case the global information, taken in its context, is meaningful, but the information associated to single bits of the signal, if processed unrelated to the context, is unreliable (Douglas et al., 1995).

In the neural approach, state variables, computation, and signal restoration are melded together, being intrinsic to the physical medium that is performing the computation (Dou-

glas et al., 1994).

Another constraint that characterizes both neural and analog implementations, and has led to the implementation of a successful and efficient strategy, is the problem of communication. While the neurally inspired analog *computation* considered here is based on principles that make it robust against noise and inhomogeneities (Douglas et al., 1995), analog *communication* is prone to noise degradation and mismatch effects (Douglas et al., 1994; Murray et al., 1991). Digital communication, on the other hand, is robust and less sensitive to noise. The strategy adopted by real neurons is to use a hybrid approach, exploiting analog computation and digital communication (Murray et al., 1991): in the dendrites and the soma, computation is performed on analog signals derived from the continuous transformation of "digital" all–or–none pulses into analog currents by the synapses; The "result" of computation is then transformed back to pulses in the axon hillock, and transmitted along axons to the synapses which transfer information to other neurons. In neuromorphic chips we use the same strategy: instead of transmitting the analog state of the neurons, we transmit the spiking activity of each neuron as a sequence of asynchronous digital events. Analog information is self–encoded in the timing of the digital events (Lazzaro et al., 1993).

The neuromorphic approach takes into account and exploits the differences between wet–ware and analog silicon. Some of them regard limitations of the silicon medium in comparison with the neural substrate, such as the lack of a full tridimensional space for wiring, or the limited fan–in and fan–out of silicon circuits (Maher et al., 1989) compared with the vast dendritic and axonal arborization of cortical cells, which can transmit signals to up to 10^5 synaptic connections (Laughlin and Sejnowski, 2003).

Neural "technology" also has limitations against silicon implementations, like the lack of low resistance wiring, and speed. Some of the advantages of silicon can be used to work around its limitations, for example the high speed of transmission can be exploited for multiplexing signals on the limited number of connections that can be designed on a silicon wafer. Neuromorphic analog chips are characterized by high parallelism, many computational units, and high connectivity; when connecting neurons between different chips, a strict limiting factor is the number of output pads (a typical chip package has 120 pins). This limit prevents the implementation of direct connections between the neurons, and requires a time multiplexing strategy (Maher et al., 1989): all neurons share the same bus to transmit their pulses, together with the (implicit) timing information. The identity, or address, of the neuron that produces the event is transmitted on the bus, from here the name "Address–Event–Representation" (AER) of the communication protocol commonly used in neuromorphic systems (Lazzaro et al., 1993).

Real neurons typically respond to novelties in the stimulus: various types of adaptation enhance time variations, and local interactions enhance spatial discontinuities (e.g. edges) in topological maps (Kiper and Carandini, 2003; Hubel and Wiesel, 1962). Activity in neural systems is event–driven: it is triggered only when and where there is a change in the stimulus, resulting in sparse signals in both time and space. This type of computation is efficient and produces communication which is intrinsically energy efficient (Lazzaro et al., 1993; Boahen, 2000). The natural approach of neuromorphic hardware is to exploit the same strategy of neural systems, implementing an asynchronous event–driven protocol: instead of scanning the whole neuronal array and synchronously transmitting information about the whole array, including both active and inactive states, the neuron itself sends its own spikes to the external bus when they are produced; the bus is occupied only when it is needed,

and its use varies linearly with the number of active neurons — usually less than the whole array. Bus occupancy is therefore optimized in terms of energy efficiency (Lazzaro et al., 1993).

By being aware of the similarities and differences between the two systems, silicon emulation can maintain its relevance in the study of brain computation.

The study of neural computation in the context of its physical primitives and constraints could in principle be simulated on general purpose digital computers, which have the advantage of high flexibility and programmability over the silicon implementation. However there are two main limitations to this approach.

From the computational point of view, a simulation of a system interacting with the real world requires the explicit mathematical description of any process, comprising model, environment, noise, and also time, that have to be explicitly encoded (Douglas et al., 1995).

From the simulation point of view, neural models comprise many non–linear interactions and couplings; The inclusion of learning and spike–based neural models increases the time scale range, resulting in stiff differential equations that are slow to solve. Additionally, the intrinsically serial architectures behind digital simulations scales the simulation time with the number of elements and couplings in the network.

Silicon emulations, on the contrary, operate in real–time, independent of the number of elements or couplings in the networks. They are effective when employed for computationally intensive problems with highly non–linear dynamics that repeat over time, where the speed of processing can in principle outperform software simulations.

This approach therefore leads to the design of devices and systems that are more efficient than digital computers in tasks requiring interaction with the real world in real–time.

We can subdivide neuromorphic engineering research into three main levels. The first level deals with the modeling of the constitutive elements of neural systems, such as neurons (Mahowald and Douglas, 1991; Indiveri, 2003b), synapses (Chicca et al., 2003b; Bartolozzi and Indiveri, 2007), photoreceptors (Delbrück and Mead, 1995), etc.

There is a trade–off between the details incorporated in such circuits, and the area they occupy on the silicon wafer. We can take as example the design of silicon neurons. One approach, adopted by Mahowald and Douglas (1991) and Rasche et al. (1997), is to realize a biophysically realistic implementation of neurons: it consists in exploiting the channel conductance of transistors to model accurately the passive and active ionic conductances of neuronal membranes, and use such circuits to implement an approximation of the Hodgkin and Huxley (1952) model. Mahowald and Douglas (1991)'s silicon neuron faithfully reproduces the behavior of real neurons and allows the characteristics of the circuit to be tuned to model different types of neurons. It models in detail action potential generation thanks to sub–circuits modeling voltage–dependent sodium conductance activation and inactivation, and delayed potassium conductance activation. Additional circuits can be used to model conductances dependent on intracellular calcium concentration, proportional to the neuron's recent spiking activity, and to implement the mechanism of spike–frequency adaptation. This approach has the major drawback of requiring large silicon area and many voltage biases to tune the behavior of the neuron, and is not suitable for the implementation of dense arrays of neurons on single chips. A different approach (Mead, 1989; Boa-

hen, 1997; van Schaik, 2001; Indiveri, 2003b) consists in phenomenologically modeling the characteristics of real neurons. This approach achieves a good approximation of the neuron's properties, including adaptation, while maintaining compactness and low–power consumption, required for the implementation of dense neural arrays.

The second level of neuromorphic engineering research focuses on the implementation of networks comprising the above mentioned elementary units, and the study of their computational performance (Chicca, 2006; Indiveri, 2002; Hahnloser et al., 2000).

At both levels the focus of research is the study of the mechanisms that render the systems adaptive to the environment, explicitly studying learning, connectivity, and relative computation (Indiveri et al., 2006; Boahen, 2005).

Advances in this field lead naturally to the third level of neuromorphic engineering research: the realization of systems with specific practical applications. Specifically, up to now, applications have been oriented toward perceptive modules such as retinae and cochleas, since they represent the front–end interface for transforming signals from the real–world into appropriate signals for connecting to computational devices for further processing. Their biological counterparts have been thoroughly studied for the same reason, and their functionality has reached a deep enough level of understanding for engineers to transpose such knowledge onto silicon. These devices are currently developed with the goal of implementing a feasible alternative to classical digital devices, especially for the realization of autonomous implantable prostheses. In this regard, they have the benefit of producing signals that are intrinsically linked to the real biological signals, and therefore can be better interfaced to healthy wet–ware.

Recently, Sarpeshkar (2006) implemented a silicon cochlea that can substitute conventional cochlear prostheses. His processor consumes such a small amount of power that it can be used for about 30 years before a surgical operation is required to change the battery. Preliminary tests on a deaf woman revealed that the processor conveyed a good signal to the auditory nerve, allowing good speech recognition in the higher cortical areas.

Lichtsteiner et al. (2006a) designed a silicon transient imager, which is used in various fields, from research in fruit flies for monitoring their wing beats, to the practical application of monitoring high–way cars for traffic control.

Zaghloul and Boahen (2006) designed a silicon retina, that, together with photo–transduction functionality, reproduces the computation performed by the cells in the retinal layers. Its output corresponds to the output of ganglion cells and could be conveyed directly to the optic nerve, realizing an implantable retinal prosthesis with high computational power and very low power consumption.

In this context I developed an additional device, the Selective–Attention Chip (SAC), that goes beyond the modeling of front–end signal processing. The chip I describe in this thesis comprises the constitutive elements of neural systems, such as neurons and synapses, organized to perform a type of cooperative/competitive computation. It implements a post–processing stage on sensory data received from neuromorphic sensory devices, via the AER communication system. Specifically, the SAC is capable of sequentially selecting the most active regions of its input. When a map of saliency is supplied as input to the SAC, the chip scans the sensory data in order of diminishing saliency, implementing a selective attention mechanism.

In the remaining of the introduction I describe the rationale for implementing selective attention in artificial perceptive systems, by describing its role in biological systems and reviewing the basic concepts and point of views about this topic. Eventually I describe the model that I chose to implement on the SAC.

1.2 Selective Attention

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 dynamic filter that selects the most salient regions of the input, sequentially allocating computational resources, for analyzing the target's details. This serial strategy limits the computational demand with respect to full parallel processing. When attention is deployed to a certain region or feature of the input space, the corresponding cortical representation is enhanced. Neuroimaging studies (see Pessoa et al. (2003) for a review) and extracellular recordings in monkey visual cortex (area MT and V4) show enhancement of the representation of the attended feature or location (Martinez-Trujillo and Treue, 2004; Reynolds et al., 2000; McAdams and Maunsell, 2000; Reynolds and Chelazzi, 2004). Psychophysical studies (Lee et al., 1999) have shown that detection thresholds and the speed of behavioral responses are enhanced by attention; their observations are consistent with a model where the effect of attention is to activate competition among visual filters, improving the capacity of the cortex to select and process relevant information from cluttered background and noisy data (Itti et al., 2001).

Part of the research in the context of attention has been devoted to discovering the mechanisms for the selection of the attentioned target. One influential work in this field is the Feature Integration Theory of Attention, proposed by Treisman and Gelade (1980). They studied attention in the context of visual search distinguished two categories of stimuli: those that "pop–out", i.e. are immediately spotted by the observer no matter the number of distracters in the search display, and those stimuli for which an explicit search through the display items is required and for which the search time depends on the number of distracters. From these observations the authors distinguished two modes of attention, one pre–attentive, exogenous, driven by the stimulus characteristics, one attentive, endogenous, voluntarily driven by the subject on the grounds of the ongoing task. Evidence suggests that there is not a clear cut between such bottom–up and top–down attentional modes in visual search, rather there is a continuum of search difficulties where the two modes interact; specifically, stimulus driven selection depends on the difference between target and distracters, and on the similarity of the distracters (Itti and Koch, 2000; Wolfe and Horowitz, 2004). These observations confirm again that computation in the brain is not absolute, rather it depends on relative context. It is difficult to disentangle the bottom–up and top–down contributions to attentional selection, since these two pathways interact to eventually determine the *"saliency"* of stimulus, which depends on both its physical and semantic characteristics and on their relevance to the current task of the subject.

There is evidence that in some areas of the brain involved in higher order visual processing and guidance of eye movements, activity is related to the saliency of the stimuli and to attentional selection. In the frontal eye field (FEF)(Thompson et al., 2005b,a) and lateral–intraparietal (LIP) area (Colby and Goldberg, 1999; Iapata et al., 2006) of the monkey, there are topographical maps that encode for stimulus saliency; their activity results from the integration of stimulus driven selection (bottom–up) and task related modulation (top–down).

Several computational models of selective attention are based on the concept of such a *"saliency map"* (Koch and Ullman, 1985; Itti and Koch, 2000; Findlay and Walker, 1999; Wolfe, 1994) — a topographic map where activity encodes for the salience of the corresponding location in the input stimulus, irrespective of the feature that determined the saliency. A scan of the saliency map in order of decreasing salience determines the shifts of the focus of attention, and in the case of ocular movements determines the end point

of saccades used to foveate the selected target, allowing more detailed processing of the stimulus.

Some models deny the existence or need for an explicit saliency map, salience being an emerging property of the activity of many neuronal populations (Jagadeesh et al., 2001). The most influential model is the *"Biased Competition Model"* (Luck et al., 1997; Reynolds et al., 1999; Kastner and Ungerleider, 2000; Deco and Lee, 2002). The foundation of this work lies on the concept of competition of the sensory stimuli for computational resources. An interaction between the bottom–up competition and top–down modulation of the competition leads to the emergence of the attended stimulus, and consequently the enhancement of its cortical representation.

The debate on these models is still going on, inspiring new and hopefully more conclusive experiments. Both points of view are worth exploring with software simulations and hardware emulations. From the operative point of view, models based on the saliency map allow the existence of an unique, unambiguous read out, the maximum of the saliency map, which can be easily used for the control of actuators. The focus of these models is on the stimulus driven computation that generates the saliency map, and on the mechanisms for creating the attentional scan path from the map itself.

Saliency–map based models

The saliency map model proposed by Itti and Koch (2001) which implements stimulus driven visual selective attention, accounts for many psychophysical and neurophysiological observations and has features that could be used in practical applications. Fig. 1.2 shows schematically the components of the system: various independent channels extract information from the input stimulus concerning color, orientation, motion, intensity, etc, in a center surround fashion and at various spatial scales. This implementation is strongly related to the competitive mechanisms postulated in visual cortex, and again relates computation to the context of the information rather than to its absolute value. The result is a topological feature map for each channel, encoding the relative strength of the corresponding feature in each point of the visual field. The combination of feature maps gives rise to the final saliency map, where the activity of each point encodes the saliency of the corresponding region in the visual field, independently of the feature or feature combination that contributed to its saliency. The most salient locations correspond to the regions where activity from many different feature maps coincide, or locations where activity from a preferentially weighted map occurs. A Winner–Take–All (WTA) competition selects the most salient location in the saliency map, guiding the center of the attentional spot–light. A self–inhibitory mechanism deselects the current winner to allow the selection of the second most salient region of the visual field. The iteration of this selection–inhibition cycle induces scanning of the visual field in order of decreasing saliency, namely the scan–path of selective attention. The number of regions included in the scan–path depends on the duration of the inhibition that prevents reselection of the most salient stimuli, a phenomenon observed in psychophysical experiments (Posner, 1980) on visual search and named *inhibition of return* (IOR).

Itti and Koch (2001) proposed a model and its software implementation, comprising the front–end data acquisition, hierarchical computation of the saliency map, and the map scanning mechanism.

The comparison between psychophysics experiments and model predictions correlate well (Parkhurst and Niebur, 2004), even though for static images the correlations between the saliency predicted by the model and the attentional scan–path of human observers are not conclusive about the role of bottom–up effects on attentional guidance (Einhäuser and

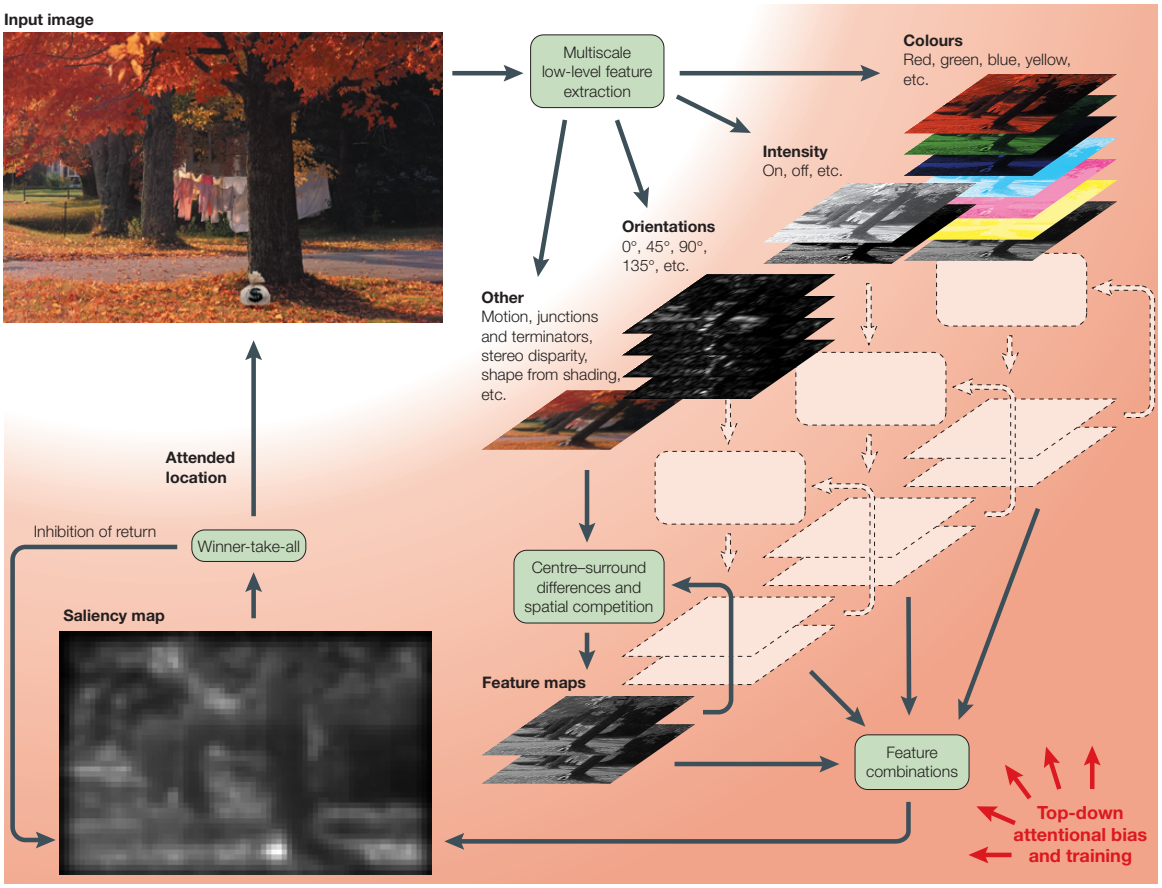


Figure 1.2: Bottom–up model of selective attention, adapted from Itti and Koch (2001). The input image is processed via different parallel independent channels, that generate topological feature maps; the combination of these feature maps creates the saliency map, that encodes for the local conspicuity of each pixel of the input image, independently of the features that generated its salience; a WTA competition selects the most salient pixel of the image, a IOR mechanism deselects the current winner, creating the attentional scan–path.

König, 2003). New studies (Itti, 2005a; Carmi and Itti, 2006) show a causal relationship between bottom–up features and attentional shifts, especially when the predictive feature is motion.

The Selective–Attention Chip (SAC) described in this thesis can be used to implement selective attention systems of the type proposed by Itti and Koch (2000) in hardware. It is designed to receive a saliency map as input and generate the attentional scan path as output, by implementing WTA and IOR mechanisms. In particular, the SAC was designed to be selective to changes of the input stimuli, which are strong predictors of attentional selection (Itti, 2005a; Carmi and Itti, 2006). Additionally the IOR mechanism can be disabled and additional instances of the same chip can be used to implement the pre–processing competitive stages for the creation of feature maps.

The SAC is therefore a useful tool for investigating different models for the bottom–up attentive system, as well as being a useful building block for the design of artificial systems that can interact with the real world in real time.

As to demonstrate its potential use in practical selective attention systems, I used the SAC and a silicon transient imager (Lichtsteiner and Delbrück, 2005) to build a two–chip visual selective attention system, and observed its response properties both with and without sensor movements. The rationale for studying the system behavior in both conditions

arises from the observation of psychophysics experiments performed for the study of visual attention. In general they can be divided into two broad categories based on whether or not eye movements for orienting towards the attentional target are allowed.

In one case, the subject is asked to maintain the eyes focused on a fixation point, *covertly* deploying attention. In these types of experiments the presence of attention is inferred from the enhancement of processing of visual stimuli, by reduction of reaction time or decrease in detection or discrimination thresholds (Lee et al., 1999).

In the other case, the subject performs ballistic eye movements, the saccades, to bring the location selected as attentional target to the fovea, where the resolution of the retina is maximal. In such a case, attention is *overtly* deployed.

Evidence that covert attention is deployed at the location of the end–point of the saccadic movement, before the movement is performed (Hoffman and Subramaniam, 1995), supports the use of eye movements as indicators for attentional selection in free vision (Findlay, 2005). The link between covert and overt attention is also supported by recordings in the Frontal Eye Fields of the monkey (Bichot and Schall, 2005): visually driven neurons in this area respond to the saliency of stimuli, integrating visual attributes and goal–driven modulation; these neurons predict the target location of saccades, while another set of movement–related neurons predict if and when a saccade will be executed, overtly pushing attention to the targets selected by the visually–driven neurons.

In this thesis I describe the experiments carried out with the two–chip selective attention system, where I used both covert and overt approaches to study the focus of attention scan–path properties, and to verify the system's functionality when it is required to perform actions in response to real world stimulation.

1.3 Thesis Outline

The work presented in this thesis includes all of the three levels in which I divided neuromorphic hardware design research, ranging from the design of a silicon synapse to the realization of a multi–chip system system implementing a model of selective attention.

In Chap. 2 I describe a novel synaptic circuit, designed to support the elaborate dynamic mechanisms observed in selective attention systems. This silicon synapse reproduces the currents originated by presynaptic action potentials across the postsynaptic membrane. The time course of such currents can be modeled with exponentials (Destexhe et al., 1998), reproduced on silicon by exploiting the voltage–current exponential transfer function of subthreshold CMOS. Additional dynamic circuits, extending the functionality of the proposed synapse, enrich the synaptic primitives of neuromorphic systems, allowing for the exploration of their computational role.

The proposed circuit is compatible with existing circuits which implement learning, short–term adaptation, and global synaptic scaling for implementing homeostatic plasticity (Turrigiano et al., 1998). The last property adapts the computational substrate to changes of the environment on very long time scales and is a biomimetic strategy for reducing inhomogeneities between neurons, becoming one of such adaptive mechanisms postulated by Mead for obtaining robust and precise computation in face of inhomogeneous and noisy constitutive elements.

In Chap. 3 I describe the circuit implementation of a classical functionality observed in neural networks required to model selective attention: the Winner–Take–All (WTA) circuit. WTA is an emergent computational property of recurrently connected neurons, which enhances the activity of neurons receiving the strongest input and suppresses the activity of

neurons receiving weaker input signals. It is one of the competitive–cooperative computational strategies capable of extracting information from noisy and ambiguous data (Douglas et al., 1995, 1999; Chicca, 2006).

In Chap. 4 I describe the Selective–Attention Chip (SAC), which comprises synapses, neurons, and the WTA circuit. The SAC extends the study of neuromorphic implementation of perceptive systems to a further stage of the processing hierarchy, going beyond data acquisition and the processing performed by silicon retinae and cochleae developed up to now (Sarpeshkar, 2006; Zaghloul and Boahen, 2006; Lichtsteiner et al., 2006b; Chan et al., 2006).

The SAC is the evolution of previously proposed selective attention chips (Indiveri, 2001b); besides incremental improvements to the circuits and the introduction of new synaptic circuits, the chip I developed includes adaptive properties in the synapses and in the neurons — Short–Term Depression (Rasche and Hahnloser, 2001) and spike frequency adaptation (Indiveri, 2003b) respectively — that I show to play a crucial role in making the chip sensitive to variations of the input stimuli, and in reducing the amount of information sent to the output bus.

Motivated by the need for building useful devices for practical applications, in Chap. 5 I describe possible uses of the SAC, in particular by building a multi–chip system with the SAC connected to a silicon retina mounted on an actuator. I use this multi–chip system to validate the use of the SAC as a tool for testing different models of selective attention, and for studying the effects of the newly included adaptive properties and of the different parameters of the networks.

Finally, in Chap. 6 I discuss the relevance of the work described in this thesis in the context of neuromorphic engineering research, and in the context of the implementation of selective attention systems.

Chapter 2

Analog VLSI synapse circuits

2.1 Introduction

Synapses are highly specialized structures which, by means of complex chemical reactions, allow neurons to transmit signals to other neurons. When an action potential generated by a neuron reaches a presynaptic terminal, a cascade of events leads to the release of neurotransmitters that give rise to a flow of ionic currents into or out of the postsynaptic neuron's membrane. These excitatory or inhibitory postsynaptic currents (EPSCs or IPSCs respectively) have temporal dynamics with a characteristic time course that can last several hundred milliseconds (Koch, 1999b).

In computational models of neural systems the temporal dynamics of synaptic currents have often been neglected. In models that represent information with mean firing rates, synaptic transmission is typically modeled as an instantaneous multiplier operator (Hertz et al., 1991). Similarly in pulse–based neural models, where the precise timing of spikes and the dynamics of the neuron's transfer function play an important role, synaptic currents are often reduced to simple instantaneous charge impulses. In VLSI implementations of neural systems, silicon synapses have also often been reduced to either simple multiplier circuits (Borgstrom et al., 1990; Satyanarayana et al., 1992), or constant current sources activated only for the duration of the presynaptic input pulse (Mead, 1989; Fusi et al., 2000; Chicca et al., 2003a).

In the context of pulse–based neural networks, modeling the detailed dynamics of post-synaptic currents can be a crucial step for learning neural codes and encoding spatio–temporal patterns of spikes. Leaky integrate–and–fire (I&F) neurons can distinguish between different temporal input spike patterns only if the synapses stimulated by the input spike patterns exhibit dynamics with time constants comparable to the time constant of the neuron's membrane potential (Gütig and Sompolinsky, 2006).

Modeling the temporal dynamics of each synapse in a network of I&F neurons can be very onerous in terms of CPU usage for software simulations, and in terms of silicon real–estate for dedicated VLSI implementations. A compromise between highly detailed models of synaptic dynamics and no dynamics at all, is to use computationally efficient models that account for the basic properties of synaptic transmission. A very efficient model that reproduces the macroscopic properties of synaptic transmission and accounts for the linear summation property of postsynaptic currents is the one based on pure exponentials proposed by Destexhe et al. (1998). In this chapter I describe a new VLSI synaptic circuit, the diff–pair integrator (DPI), that implements the model proposed in Destexhe et al. (1998) as a log–domain linear temporal filter, and that supports a wide range of synaptic properties ranging from short–term depression to conductance–based EPSC generation.

2.1.1 Biological synapses

Synapses are the connections between neurons, implementing information transfer from the presynaptic to the postsynaptic neuron. They can be divided into two main categories, electric and chemical; I focus on the second type, as they are basic building blocks of computational and circuital neural models. Fig. 2.1 schematically illustrates the chain of events that take place during synaptic transmission: a presynaptic neuron produces action potentials that are related to its state; these stereotyped all–or–none events travel along the axon reaching synaptic boutons, specialized structures of the axon that connect to the dendritic arbor of the postsynaptic neurons. The action potential causes a depolarization of the bouton membrane and the influx of calcium into the terminal, which in turns causes the release of vesicles into the space between the presynaptic bouton and the postsynaptic membrane, the synaptic cleft. Each vesicle contains a quantum of neurotransmitters, which bind to specialized receptors of the postsynaptic neuron. The receptors can be ionic channels that open at the binding of the neurotransmitter (ionotropic channels), or can cause opening of channels via an intracellular second messenger (metabotropic). Channels are transmembrane proteins that change their conformation to allow the flow of specific ions through the postsynaptic membrane. Their opening modifies the membrane conductance g_{ion} of their particular ion, causing a current flux $I_{ion} = g_{ion}(V_{mem} - E_{ion})$, proportional to the difference between the membrane voltage and the reversal potential of the ion, determined by the different ionic concentration across the membrane. The variation in conductance has a typical time course that depends on the dynamics of the neurotransmitter binding and un–binding from the receptor, and results in the typical shape of the postsynaptic currents. In the next paragraph I show how these currents have time courses that can be modeled using exponentials. These currents in turn modify, either depolarizing (EPSCs, excitatory postsynaptic currents), or hyperpolarizing (IPSCs, inhibitory postsynaptic currents), the postsynaptic membrane potential. Examples of receptors that cause EPSCs are the *ligand–gated ionotropic* AMPA channels, which let Na^+ and K^+ ions flow into and out of the postsynaptic neuron respectively, when they bind glutamate released by the presynaptic terminal. $GABA_\alpha$ receptors instead generate inhibitory chlorid influx into the postsynaptic neuron. Other channels need the simultaneous presence of the neurotransmitter and sufficient depolarization of the postsynaptic membrane to open and change the membrane conductance to their ion, this is the case of the *ligand–* and *voltage–gated* NMDA receptors. The net contribution of many of the synaptic currents from the whole dendritic arbor can cause sufficient depolarization of the postsynaptic membrane to reach the axon–hillock, where an action potential can be generated. All of the events involved in synaptic transmission, from the release of vesicles to the ion flux, are generated by molecular processes with many different contributions and possible modulations, which renders synaptic transmission not just a simple connection, but a site of information processing and computation. In the following parts of this chapter I describe the various circuits developed in the past 20 years that precede the one I propose, and present additional circuits that enrich the behavior of silicon synaptic transmission.

2.1.2 Computational models of biological synapses

In biological synapses, neurotransmitter release from a single release site is probabilistic. In the design of silicon synapses it is commonly assumed that an ensemble of release sites has a mean behavior that is deterministic.

The time evolution of EPSCs recorded from biological synapses are typically fitted by

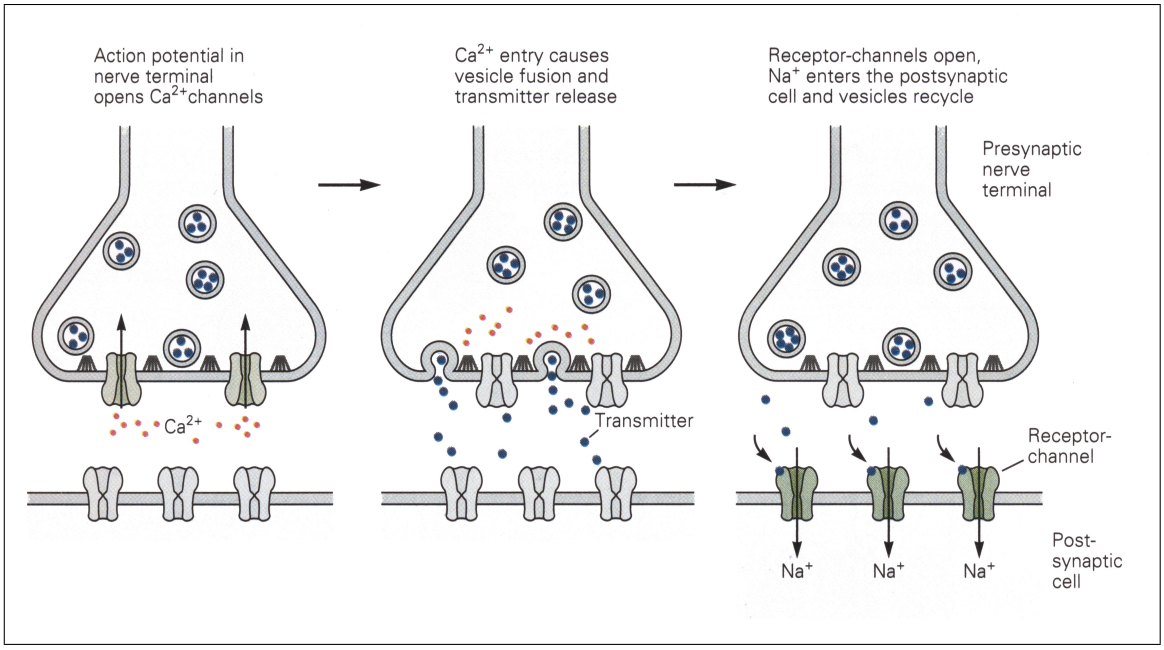


Figure 2.1: Schematic illustration of the events involved in the synaptic transmission. Adapted from Kandel et al. (2000).

α–functions[1] (Rall, 1967; Koch, 1999a). However Destexhe et al. (1998) show that post-synaptic current dynamics can be faithfully modeled by pure exponentials, in a framework where synaptic reactions are described with kinetic equations.

Such equations describe the kinematics of binding and un–binding of neurotransmitters on postsynaptic receptors that cause the corresponding proteic channels to open and let ionic currents flow through the postsynaptic neuron's membrane. For a large class of synapses, named *ligand–gated*, opening of channels depends only on the binding of the neurotransmitters that directly cause the current influx. In this case the activation and inactivation of membrane channels is well described by the reaction

$$R + N \underset{\beta}{\overset{\alpha}{\rightleftarrows}} NR*,$$

where α and β are the voltage independent forward and backward rate constants, N is the neurotransmitter concentration and R are the postsynaptic receptors. Assuming that the change in neurotransmitter concentration $[N]$ in the cleft occurs in a brief pulse, and defining r as the fraction of receptors in the activated state, the kinetic model can be translated to the first–order differential equation

$$\frac{dr}{dt} = \alpha \cdot [N](1 - r) - \beta r.$$

The current flowing into the neuron is then: $I_{syn} = \overline{g} \cdot r\,(V - E_{syn})$. Solving the differential equation leads to the expression of r for the two conditions during (eq. (2.1)) and after (eq. (2.2)) the pulse of neurotransmitter concentration $[N]$ as

$$r = \left(r_0 - \frac{\alpha \cdot [N]}{\alpha \cdot [N] + \beta} \right) e^{-(\alpha \cdot [N] + \beta)(t - t_0)} + \frac{\alpha \cdot [N]}{\alpha \cdot [N] + \beta}, \tag{2.1}$$

[1] $g(t) = \alpha T e^{-\alpha T}$, where $\alpha = \frac{\tau_m}{t_{peak}}$ and $T = \frac{t}{\tau_m}$; τ_m is the membrane time constant

$$r = r_0 e^{-\beta(t-t_0)}. \tag{2.2}$$

In Sec 2.2.6 I propose a neuromorphic synapse circuit (the DPI) and demonstrate that its analytical solution circuit leads to the same exact time dependence of the postsynaptic currents, both for the charging and the discharging phase.

2.1.3 Neuromorphic synapse emulation

Synaptic circuits translate presynaptic voltage pulses into postsynaptic currents injected in the membrane of their target neuron, with a gain typically referred to as the *synaptic weight*. The function of translating "fast" presynaptic pulses into long–lasting postsynaptic currents, with elaborate temporal dynamics, can be efficiently mapped onto silicon using subthreshold (or *weak–inversion*) analog VLSI (aVLSI) circuits (Liu et al., 2002). In typical VLSI neural network architectures, the currents generated by multiple synapses are integrated by one single postsynaptic neuron circuit. The neuron circuit carries out a weighted sum of the input signals, produces postsynaptic potentials and eventually generates output spikes, which are typically transmitted to synaptic circuits in further processing stages. A very common neuron model used in VLSI spike–based neural networks is the *point–neuron*. With this model the spatial position of the synaptic circuits connected to the neuron is not relevant and the currents produced by the synapses are summed linearly into the single neuron's membrane capacitance node. Alternatively, synaptic circuits (including the one presented in this thesis) can be integrated in multi–compartmental models of neurons, and the neuron's dendrite, comprising the spatial arrangement of VLSI synapses connected to the neuron, implements the spatial summation of synaptic currents (Northmore and Elias, 1998; Arthur and Boahen, 2004).

Irrespective of the neuron model used, one of the main requirements for synaptic circuits in large VLSI neural networks is compactness: the less silicon area is used, the more synapses can be integrated on the chip. On the other hand, implementing synaptic integrator circuits with linear response properties and time constants of the order of tens of milliseconds can require substantial silicon area. Therefore designing VLSI synaptic circuits that are compact and linear, and that model relevant functional properties of biological synapses, is a challenging task still being actively pursued. Several subthreshold synaptic circuit designs have been proposed in the past (Mead, 1989; Lazzaro, 1994; Boahen, 1998; Fusi et al., 2000; Chicca et al., 2003b; Shi and Horiuchi, 2004a; Gordon et al., 2004; Hynna and Boahen, 2006) covering a range of trade–offs between functionality and complexity of temporal dynamics versus circuit and layout size. Some of the proposed circuits require floating–gate devices (Gordon et al., 2004) or restrict the input or output signals to a very limited dynamic range (Hynna and Boahen, 2006) to reproduce in great detail the physics of biological synaptic channels.

As discussed in Sec. 2.1.2, phenomenological models reproducing the time course of real EPSCs work as well, showing a wide range of properties such as temporal summation of synaptic currents; at the same time their hardware implementation is more feasible. The design of the diff–pair integrator (DPI) synapse proposed in this thesis is inspired by a series of functionally equivalent circuits proposed in the literature that implement kinetic models of synaptic transmission; these circuits collectively share many of the advantages of the DPI, but individually lack one or more of the features of our design.

In the next section I present an overview of previously proposed synaptic circuits, and describe the DPI synapse pointing out the advantages that the DPI offers over each of them.

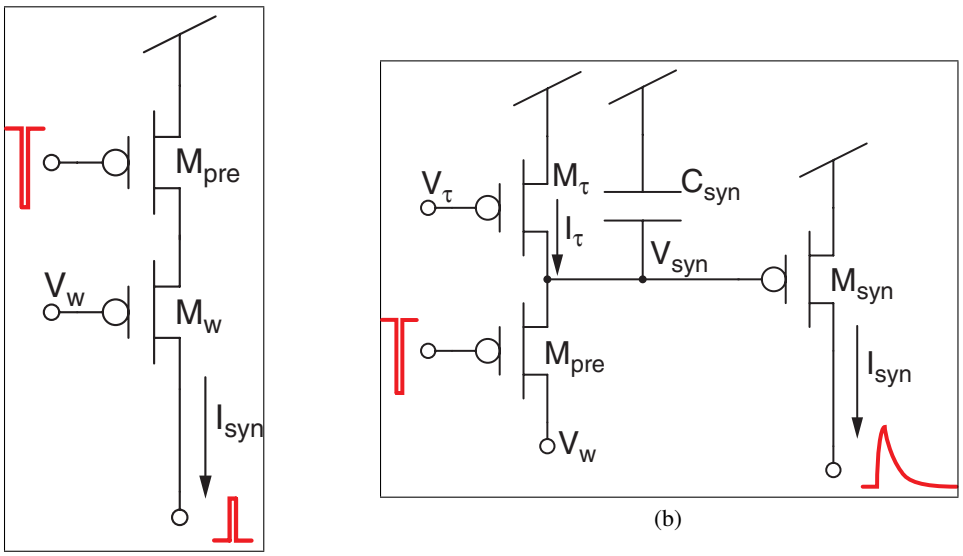


Figure 2.2: (a) Pulsed current–source synaptic circuit. (b) Reset–and–discharge synapse.

2.2 State–of–the–art and Diff–Pair Integrator synapse

2.2.1 Pulsed current–source synapse

The pulsed current–source synapse, originally proposed by Mead (1989) in the late 80s, was one of the first synaptic circuits implemented using transistors operated in the subthreshold domain. The circuit schematics are shown in Fig. 2.2(a): it consists of a voltage controlled current–source activated by an active–low input spike. In VLSI pulsed–neural networks, input spikes are typically brief digital voltage pulses that last at most a few micro–seconds. The output of this circuit is a pulsed current I_{syn} that lasts as long as the input spike. Assuming that the output p–FET M_w is saturated (*i.e.* that its V_{ds} is greater than $4U_T$), the current I_{syn} can be expressed as

$$I_{syn} = I_0 e^{-\frac{\kappa}{U_T}(V_w - V_{dd})}, \tag{2.3}$$

where V_{dd} is the power supply voltage, I_0 the leakage current, κ is the subthreshold slope factor, and U_T is the thermal voltage (Liu et al., 2002).

This circuit is extremely compact, but does not integrate input spikes into continuous output currents. Whenever a presynaptic spike reaches M_{pre}, the postsynaptic membrane potential undergoes a step increase proportional to I_{syn}. As integration only happens at the level of the postsynaptic I&F neuron, input spike trains with same mean rates but with different spike timing distributions cannot be distinguished. However, given its simplicity and compactness, this circuit has been used in a wide variety of VLSI implementations of pulse–based neural networks that use mean firing rates as the neural code (Murray, 1998; Fusi et al., 2000; Chicca et al., 2003a).

2.2.2 Reset–and–discharge synapse

In the early 90s, Lazzaro (1994) proposed a synaptic circuit where the duration of the output EPSC, $I_{syn}(t)$, could be extended with respect to the input voltage pulse by means of

a tunable exponential decay (see also (Shi and Horiuchi, 2004b) for a recent application example). This circuit, shown in Fig. 2.2(b), comprises three p–FET transistors and one capacitor; the p–FET M_{pre} is used as a digital switch which is turned on by the synapse's input spikes; the p–FET M_τ is operated in subthreshold and is used as a constant current-source to linearly charge the capacitor C_{syn}; the output p–FET M_{syn} is used to generate an EPSC that is exponentially dependent on the V_{syn} node (assuming subthreshold operation and saturation), the equation for which is given by

$$I_{syn}(t) = I_0 e^{-\frac{\kappa}{U_T}(V_{syn}(t)-V_{dd})}. \tag{2.4}$$

At the onset of each presynaptic pulse the node V_{syn} is (re)set to the bias V_w. When the input pulse ends, the p–FET M_{pre} is switched off and the node V_{syn} is driven linearly back to V_{dd}, at a rate set by I_τ/C_{syn}. For subthreshold values of $(V_{dd} - V_w)$, the EPSC generated by an input spike is therefore given by

$$I_{syn} = I_{w0} e^{-\frac{t}{\tau}}, \tag{2.5}$$

where $I_{w0} = I_0 e^{-\frac{\kappa}{U_T}(V_w - V_{dd})}$, and $\tau = \frac{\kappa I_\tau}{U_T C_{syn}}$.

In general, given a generic spike sequence on n spikes

$$\rho(t) = \sum_i^n \delta(t - t_i), \tag{2.6}$$

the response of the "reset–and–discharge" synapse can be formally expressed as

$$I_{syn}(t) \;=\; I_{w0} e^{-\frac{t}{\tau}} \cdot \int_0^t \delta(\xi - t_n) e^{\frac{\xi}{\tau}} d\xi \;=\; I_{w0} e^{-\frac{(t-t_n)}{\tau}}. \tag{2.7}$$

Although this synaptic circuit produces an EPSC which lasts longer than the duration of its input pulses, and which decays exponentially with time, its response depends only on the last (nth) input spike. This non–linear property of the circuit fails to reproduce the linear summation property of postsynaptic currents often desired in synaptic models, and makes the theoretical analysis of networks of neurons interconnected with this synapse intractable.

2.2.3 Linear charge–and–discharge synapse

Fig. 2.3(a) shows a modification of the reset–and–discharge synapse that has been often used by the neuromorphic engineering community, and that was recently presented in (Arthur and Boahen, 2004). Here the presynaptic pulse, applied to the input n-FET M_{pre}, is active high. Assuming that all transistors are saturated and operate in subthreshold, the circuit behavior is the following: During an input pulse, the node $V_{syn}(t)$ decreases linearly at a rate set by the net current $I_w - I_\tau$ and the synapse EPSC $I_{syn}(t)$ increases exponentially (charge phase). In between spikes, the $V_{syn}(t)$ node is re–charged toward V_{dd} at a rate set by I_τ and $I_{syn}(t)$ decreases exponentially with time (discharge phase). The circuit equations that describe this behavior are

$$I_{syn}(t) = \begin{cases} I^-_{syn} e^{+\frac{(t-t_i^-)}{\tau_c}} & \text{(charge phase)} \\ I^+_{syn} e^{-\frac{(t-t_i^+)}{\tau_d}} & \text{(discharge phase)}, \end{cases} \tag{2.8}$$

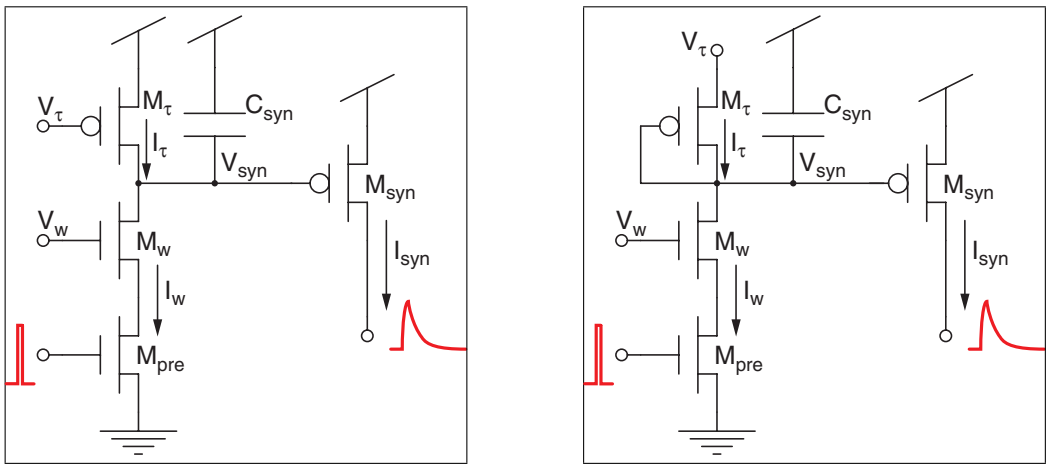


Figure 2.3: (a) Linear charge–and–discharge synapse. (b) Current mirror integrator synapse.

where t_i^- is the time at which the i^{th} input spike arrives, t_i^+ the time at which it ends, I_{syn}^- the initial condition at t_i^-, I_{syn}^+ the initial condition at t_i^+, $\tau_c = \frac{U_T C_{syn}}{\kappa(I_w - I_\tau)}$ is the charge phase time constant and $\tau_d = \frac{U_T C_{syn}}{\kappa I_\tau}$ the discharge phase time constant.

Assuming that each spike lasts a fixed brief period, Δt, and considering two successive spikes arriving at times t_i^- and t_{i+1}^-, the equation for I_{syn} is given by

$$I_{syn}(t_{i+1}^-) = I_{syn}(t_i^-) e^{\Delta t \left(\frac{1}{\tau_c} + \frac{1}{\tau_d}\right)} e^{-\frac{(t_{i+1}^- - t_i^-)}{\tau_d}}. \tag{2.9}$$

From this recursive equation one can derive the response of the linear charge–and–discharge synapse to a generic spike sequence $\rho(t)$ of n spikes as

$$I_{syn}(t) = I_0 e^{n\Delta t \left(\frac{1}{\tau_c} + \frac{1}{\tau_d}\right)} e^{-\frac{t}{\tau_d}}, \tag{2.10}$$

assuming the initial condition $V_{syn}(0) = V_{dd}$.

The EPSC dynamics depend on the total number of spikes n received at time t, and on the circuit's time constants τ_c and τ_d. By denoting the input spike train frequency at time t as $f = (n/t)$, eq. (2.10) can be re–written as

$$I_{syn}(t) = I_0 e^{-\frac{\tau_c - f\Delta t(\tau_c + \tau_d)}{\tau_c \tau_d} t}. \tag{2.11}$$

The major drawback of this circuit, aside from not being a linear integrator, is that if the argument of the exponential in eq. (2.11) is positive (*i.e.* if $f > \frac{1}{\Delta t}\frac{I_\tau}{I_w}$) the output current increases exponentially with time and the circuit's response saturates: $V_{syn}(t)$ decreases all the way to Gnd, and $I_{syn}(t)$ increases to its maximum value. This can be a problem because, in these conditions, the circuit's steady state response does not encode the input frequency.

2.2.4 Current–mirror–integrator synapse

In his PhD dissertation, Boahen (1997) proposed a synaptic circuit which differs from the linear charge–and–discharge circuit by a single node connection (see Fig. 2.3), but which

has a dramatically different behavior. The two transistors $M_\tau - M_{syn}$ of Fig. 2.3(b) implement a p–type current mirror, and form a current mirror integrator (CMI) together with the capacitor C_{syn}. The CMI synapse implements a non–linear pulse integrator circuit which produces a mean output current I_{syn} which increases with input firing rates, and which has a saturating non–linearity with a maximum amplitude which depends on the circuit's synaptic weight bias V_w and on its *time–constant*[2] bias V_τ.

The CMI response properties have been derived analytically by Hynna and Boahen (2001) for steady state conditions. An explicit solution of the CMI response to a generic spike train, that does not require the steady state assumption, was also derived by Chicca (2006). According to the analysis presented in (Chicca, 2006), the CMI response to a spike arriving at t_i^- and ending at t_i^+ is given by

$$I_{syn}(t) = \begin{cases} \dfrac{\alpha I_w}{1+\left(\frac{\alpha I_w}{I_{syn}^-}-1\right)e^{-\frac{(t-t_i^-)}{\tau_c}}} & \text{(charge phase)} \\ \dfrac{I_w}{\frac{I_w}{I_{syn}^+}+\frac{\left(t-t_i^+\right)}{\tau_d}} & \text{(discharge phase),} \end{cases} \tag{2.12}$$

where t_i^-, t_i^+, I_{syn}^-, and I_{syn}^+ are the same as defined in eq. (2.8), $\alpha = e^{\frac{(V_\tau - V_{dd})}{U_T}}$, $\tau_c = \frac{C_{syn} U_T}{\kappa I_w}$, and $\tau_d = \alpha \tau_c$.

During the charge phase the EPSC increases over time as a sigmoidal function, while during the discharge phase it decreases with a $1/t$ profile. The discharge of the EPSC is therefore extremely fast compared to the typical exponential decay profiles of other synaptic circuits. The parameter α (set by the V_τ bias voltage) can be used to slow down the EPSC response profile. However, this parameter affects both the length of the EPSC discharge profile and the maximum amplitude of the EPSC charge phase: longer response times (larger values of τ_d) produce higher EPSC values.

Despite these problems, and even though the CMI cannot be used to linearly sum post-synaptic currents, this circuit was very popular and has been extensively used by the neuromorphic engineering community in the past (Boahen, 1998; Horiuchi and Hynna, 2001; Indiveri, 2000a; Liu et al., 2001).

2.2.5 Log–domain integrator synapse

More recently Merolla and Boahen (2004) proposed another variant of the linear charge–and–discharge synapse that implements a *true* linear integrator circuit. This circuit (shown in Fig. 2.4) exploits the logarithmic relationship between subthreshold MOSFET gate–to–source voltages and their channel currents, and is therefore called a *log–domain* filter. The output current I_{syn} of this circuit has the same exponential dependence on its gate voltage V_{syn}, as all other synapses presented (see eq. (2.4)). Therefore its derivative with respect to time is

$$\frac{d}{dt} I_{syn} = -I_{syn} \frac{\kappa}{U_T} \frac{d}{dt} V_{syn}. \tag{2.13}$$

During an input spike (charge phase), the dynamics of the V_{syn} are governed by the equation $C_{syn} \frac{d}{dt} V_{syn} = -(I_w - I_\tau)$. Combining this first order differential equation with

[2]The CMI does not implement a linear integrator filter, therefore the term "time–constant" is improperly used. I use it in this context to denote a parameter which controls the temporal extension of the CMI's impulse response.

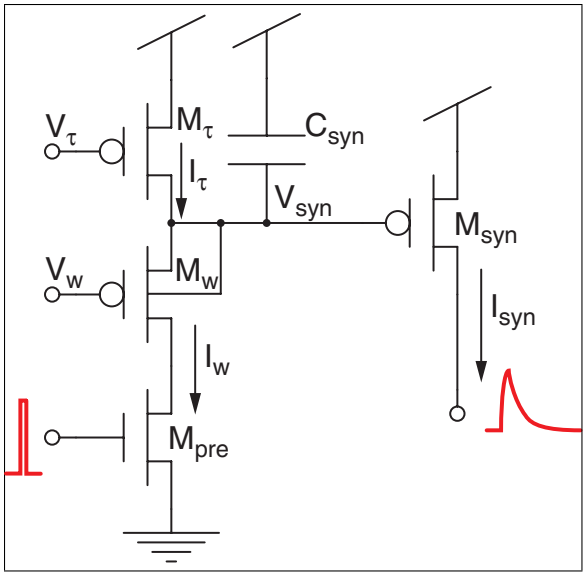


Figure 2.4: Log–domain integrator synapse.

eq. (2.13), I obtain

$$\tau \frac{d}{dt} I_{syn} + I_{syn} = I_{syn} \frac{I_w}{I_\tau}, \tag{2.14}$$

where $\tau = \frac{C_{syn} U_T}{\kappa I_\tau}$. The beauty of this circuit lies in the fact that the term I_w is inversely proportional to I_{syn} itself:

$$I_w = I_0 e^{-\frac{\kappa(V_w - V_{syn})}{U_T}} = I_0 e^{-\frac{\kappa(V_w - V_{dd})}{U_T}} e^{\frac{\kappa(V_{syn} - V_{dd})}{U_T}} = I_{w0} \frac{I_0}{I_{syn}}, \tag{2.15}$$

where I_0 is the leakage current and I_{w0} is the current flowing through M_w in the initial condition, when $V_{syn} = V_{dd}$. Substituting this expression for I_w into eq. (2.14), the right term of the differential equation loses the dependence on I_{syn} and becomes the constant factor $\frac{I_0 I_{w0}}{I_\tau}$.

Therefore the log–domain integrator transfer function takes the form of a canonical first order low–pass filter equation, and its response to a spike arriving at t_i^- and ending at t_i^+ is

$$I_{syn}(t) = \begin{cases} \frac{I_0 I_{w0}}{I_\tau} \left(1 - e^{-\frac{(t - t_i^-)}{\tau}} \right) + I_{syn}^- e^{-\frac{(t - t_i^-)}{\tau}} & \text{(charge phase)} \\ I_{syn}^+ e^{-\frac{(t - t_i^+)}{\tau}} & \text{(discharge phase)}, \end{cases} \tag{2.16}$$

This is the only synaptic circuit of the ones described up to now that has linear filtering properties. The same silicon synapse can be shared to sum the contributions of spikes potentially arriving from different sources in a linear way. This could save significant amounts of silicon real–estate in neural architectures where the synapses do not implement learning or local adaptation mechanisms, and could therefore solve many of the problems that have hindered the development of large–scale VLSI multi–neuron chips up to now. However, this particular circuit has two drawbacks. One problem is that the VLSI layout of the schematic shown in Fig. 2.4 requires more area than the layout of other synaptic circuits, because the M_w p–FET has to live in an "isolated well" structure (Liu et al., 2002). The second, and more serious problem, is that the spike lengths used in pulse–based neural network systems,

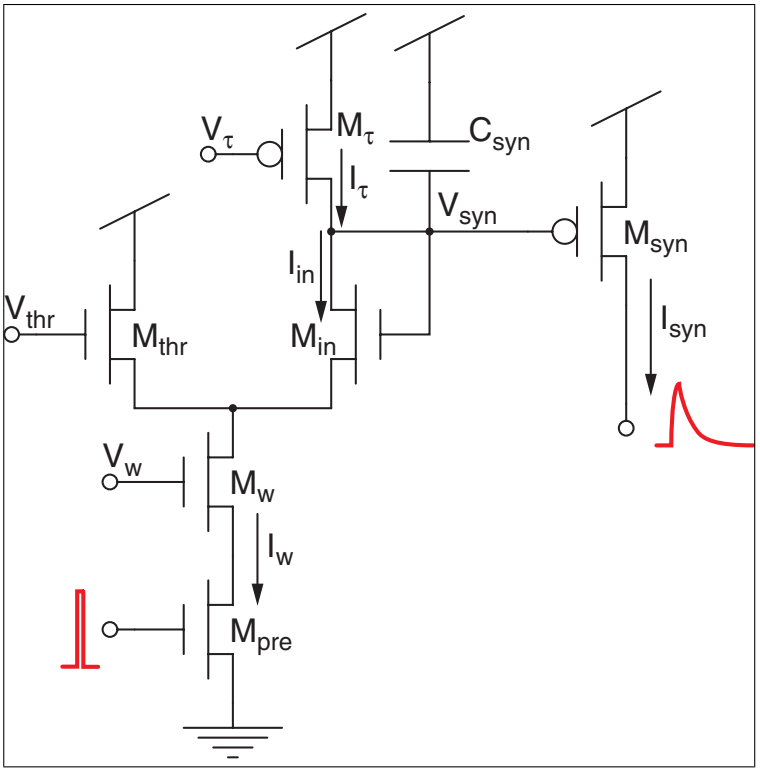


Figure 2.5: Diff–pair integrator synapse.

which typically last less than a few micro–seconds, are too short to inject enough charge into the membrane capacitor of the postsynaptic neuron to produce any effect. The maximum amount of charge possible is $\Delta Q = \frac{I_0 I_{w0}}{I_\tau}\Delta t$, and I_{w0} cannot be increased beyond subthreshold current limits (of the order of nano–amperes), or else the log–domain properties of the filter break–down (note that also I_τ is fixed, to set the desired time constant τ). A possible solution is to increase the short (off–chip) input pulse lengths with on–chip pulse extenders (*e.g.* with CMI circuits). But this solution requires additional circuitry at each input synapse, and makes the layout of the overall circuit even larger (Merolla and Boahen, 2004).

2.2.6 Diff–pair integrator synapse

The DPI circuit that I designed solves the problems of the log–domain integrator synapse while maintaining its linear filtering properties, thus preserving the possibility of multiplexing in time spikes arriving from different sources. The schematic diagram of the DPI synapse is shown in Fig. 2.5. This circuit comprises four n–FETs, two p–FETs, and a capacitor. The n–FETs form a differential pair whose branch current, I_{in}, represents the input to the synapse during the charge phase. Assuming subthreshold operation and saturation regime, and that the subthreshold slope factors of PMOS, k_p, and NMOS, k_n, are equal, the diff–pair branch current, I_{in}, can be expressed as

$$I_{in} = I_w \frac{e^{\frac{\kappa V_{syn}}{U_T}}}{e^{\frac{\kappa V_{syn}}{U_T}} + e^{\frac{\kappa V_{thr}}{U_T}}}. \tag{2.17}$$

Multiplying the numerator and denominator of eq. (2.17) by $e^{-\frac{\kappa V_{dd}}{U_T}}$, one can express I_{in} as

$$I_{in} = \frac{I_w}{1 + \left(\frac{I_{syn}}{I_{gain}}\right)}, \tag{2.18}$$

where the term $I_{gain} = I_0 e^{-\frac{\kappa(V_{thr} - V_{dd})}{U_T}}$ represents a virtual p–type subthreshold current that is not tied to any p–FET in the circuit.

As for the log–domain integrator, one can combine the C_{syn} capacitor equation $C_{syn}\frac{d}{dt}V_{syn} = -(I_{in} - I_\tau)$ with eq. (2.13), and write

$$\tau \frac{d}{dt} I_{syn} = -I_{syn}\left(1 - \frac{I_{in}}{I_\tau}\right), \tag{2.19}$$

where (as usual) $\tau = \frac{CU_T}{\kappa I_\tau}$. Replacing I_{in} from eq. (2.18) into eq. (2.19) I obtain

$$\tau \frac{d}{dt} I_{syn} + I_{syn} = \frac{I_w}{I_\tau} \frac{I_{syn}}{1 + \left(\frac{I_{syn}}{I_{gain}}\right)}. \tag{2.20}$$

This is a first order *non–linear* differential equation; however the steady–state condition can be solved in closed form, and its solution is

$$I_{syn} = \frac{I_{gain}}{I_\tau}(I_w - I_\tau). \tag{2.21}$$

If $I_w \gg I_\tau$, the output current, I_{syn}, will eventually rise to values such that $I_{syn} \gg I_{gain}$, when the circuit is stimulated with a step input signal. If $\frac{I_{syn}}{I_{gain}} \gg 1$ the I_{syn} dependence in the second term of eq. (2.20) cancels out, and the non–linear differential equation simplifies to the canonical first order low–pass filter equation

$$\tau \frac{d}{dt} I_{syn} + I_{syn} = \frac{I_w I_{gain}}{I_\tau} \tag{2.22}$$

In this case, the response of the DPI synapse to a spike arriving at t_i^- and ending at t_i^+ is

$$I_{syn}(t) = \begin{cases} \frac{I_{gain} I_w}{I_\tau}\left(1 - e^{-\frac{(t - t_i^-)}{\tau}}\right) + I_{syn}^- e^{-\frac{(t - t_i^-)}{\tau}} & \text{(charge phase)} \\ I_{syn}^+ e^{-\frac{(t - t_i^+)}{\tau}} & \text{(discharge phase),} \end{cases} \tag{2.23}$$

I can now compare the circuit's response to the synaptic computational model described in Sec. 2.1.2. For the discharge case both equations (2.2) and (2.23) (discharge phase) show the same exponential time dependence of the EPSC, with a fixed time constant that is independent of the postsynaptic membrane potential ($\beta = 1/\tau$). During the spike (or release of neurotransmitters), equations (2.1) and (2.23) (charge phase) hold. In this case I_{in} is a brief current pulse, reflecting the assumption made by Destexhe and colleagues of a brief pulse of neurotransmitter concentration $[N]$ (Destexhe et al., 1998). The general form of both equations is identical, with a difference in the exponents. This difference reflects the fact that the binding of neurotransmitters, $\alpha \cdot [N]$, depends on the fraction of receptors that

are closed, while I_{in} does not. However the decrease of I_{in} with V_{syn} can be seen as a similar mechanism, acting on the multiplicative gain factor rather than on the exponent of the exponential. The smooth saturation effect of the differential pair on I_{in} reflects the upper bound on the release of neurotransmitters from a pool of release sites, which is present in biological synapses.

The solution of the DPI synapse is almost identical to that of the log–domain integrator synapse described in eq. (2.16). The only difference is that the term I_0 of eq. (2.16) is replaced by I_{gain}. This scaling factor can be used to amplify the charge phase response amplitude, therefore solving the problem of generating sufficiently large charge packets sourced into the neuron's integrating capacitor for input spikes of very brief duration, while keeping all currents in the subthreshold regime, and without requiring additional pulse–extender circuits. In addition, the layout of the DPI synapse does not require isolated well structures, and so can be implemented in a very compact way.

Response to arbitrary spike trains

Silicon synapses are typically stimulated with trains of pulses (spikes) of very brief duration, separated by longer inter–spike intervals (ISIs). The response to such a stimulus can be written as the convolution of the impulse response of the DPI, $h(t)$, with the square function, $s(t)$, and a train of impulses, $\rho(t)$, where

$$s(t) \triangleq u(t - t_0) - u(t - t_0 - \Delta t),$$

with $u(t)$ defined as the step function, and where

$$\rho(t) = \sum_{t_i} \delta(t - t_i).$$

The response of the synapse, denoted as $g(t)$, to the train of spikes is then

$$g(t) = \rho(t) * s(t) * h(t) = \rho(t) * I_{syn}(t),$$

where $I_{syn}(t)$ is the response of the DPI to the square function, derived in eq. (2.23).

Since the system is linear, the effect of the response to each spike is independent of the response to the other spikes, and the mean response in one second corresponds to the area of the function $I_{syn}(t)$ summed as many times as the mean number of spikes in one second, i.e. the mean frequency.

The mean level of the output current of the DPI in response to a train of spikes of mean frequency, $\bar{f}$, is then

$$< I_{syn} >= \bar{f} \int_{-\infty}^{+\infty} I_{syn}(t)dt. \tag{2.24}$$

Substituting $I_{syn}(t)$ with the expression derived in eq. (2.23), and for the initial conditions $t_0 = 0$ and $I_{syn}(t_0) = 0$, the integral can be explicitly solved by

$$< I_{syn} >= \bar{f} \left[\int_0^{\Delta t} \frac{I_w I_{gain}}{I_\tau}(1 - e^{-\frac{t}{\tau}})dt + \int_{\Delta t}^{\infty} I_{syn}(\Delta t) e^{-\frac{t-\Delta t}{\tau}} dt \right]. \tag{2.25}$$

From eq. (2.23), the initial condition for the decay of the current after the spike can be derived as $I_{syn}(\Delta t) = \frac{I_w I_{gain}}{I_\tau}(1 - e^{-\frac{\Delta t}{\tau}})$. The solution of eq. (2.25) leads to the explicit

expression of the mean response of the DPI to a train of spikes, and is linear with the mean input frequency given by

$$< I_{syn} >= \frac{I_w I_{gain}}{I_\tau} \Delta t \bar{f}. \tag{2.26}$$

This property is fundamental for using the DPI to sum the activity of many different input neurons, a strategy exploited for saving silicon area when using non plastic synapses (Shi and Horiuchi, 2004a). Additionally it is crucial in applications where the output has to be linear with the input, as for the SAC implementation (see Sec. 4.2.2). The experiments performed in Sec. 4.2.2 show also that the DPI is more suitable for generating the input currents of the SAC than the CMI used in previous versions, since it generates smoother fluctuations of the current around the mean value determined by the input frequency.

For very high frequencies, the ISI becomes negligible with respect to the spike width Δt, $\bar{f} = \frac{1}{\Delta t + \overline{ISI}} \rightarrow \frac{1}{\Delta t}$ and the response saturates to $\frac{I_w I_{gain}}{I_\tau}$.

As for the log–domain integrator synapse described in Sec. 2.2.5, the DPI synapse implements a low–pass filter with a linear transfer function (under the realistic assumption that $I_w \gg I_\tau$). Although it is less compact than the synaptic circuits described in Sec. 2.2.1, 2.2.2, 2.2.3 and 2.2.4, it is the only one that can reproduce the exponential dynamics observed in excitatory and inhibitory postsynaptic currents of biological synapses (Destexhe et al., 1998), without requiring additional input pulse–extender circuits. Moreover, the DPI synapse I propose has independent control of the time constant, synaptic weight and synaptic scaling parameters. The extra degree of freedom obtained with the V_{thr} parameter can be used to globally scale the efficacies of the DPI circuits that share the same V_{thr} bias. This feature could in turn be employed to implement global homeostatic plasticity mechanisms complementary to local spike–based plasticity ones acting on the synaptic weight, V_w, node (see Sec. 2.4).

An alternative approach to those described in this section for hardware synapse emulation is based on the use of floating–gate transistors, where the floating voltage of the gate is used either to store the synaptic weight or to produce long time constants. The floating–gate approach becomes relevant when local storage and modification of the synaptic weight are required. I did not consider this approach for the design of the silicon synapse, as the selective attention project does not focus on plasticity. Nevertheless the DPI synapse is compatible with floating–gate technology for the implementation of learning algorithms.

In the next section I present experimental results from a VLSI chip comprising an array of DPI synapses connected to low–power leaky integrate and fire neurons (Indiveri et al., 2006), which validate s the analytical derivations presented here.

2.3 Experimental results

A prototype chip fabricated in standard AMS 0.35μm CMOS technology comprises the DPI circuit and additional test structures to augment the synapse's functionality. Fig. 2.6 shows a picture of the synaptic circuit layout. The full layout occupies an area of 1360μm^2; it can therefore be used to implement networks of spiking neurons with a very large number of synapses on a small chip area. For example, in a recent chip Mitra et al. (2006) implemented a network comprising 8192 DPI synapses and 32 neurons (256 synapses per neuron) using an area of only 12mm^2. The silicon area occupied by the synaptic circuit can

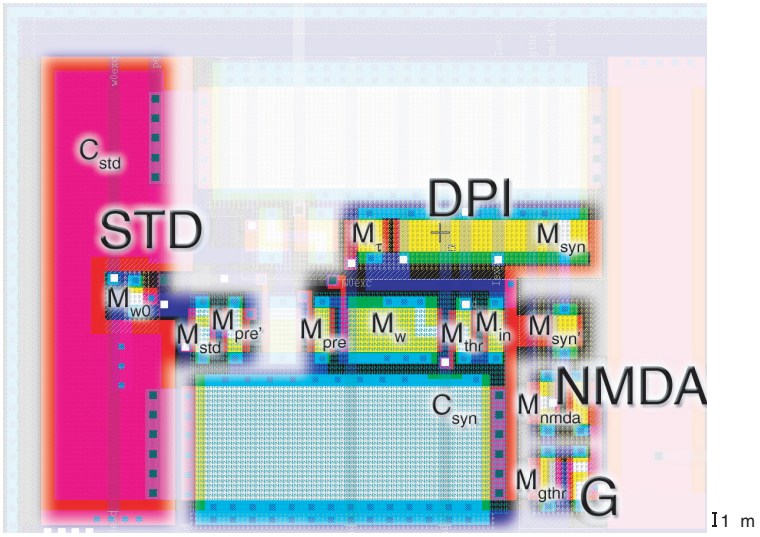


Figure 2.6: Layout of the fabricated DPI synapse and additional circuits that augment the synapse's functionality. The elements names and the STD, NMDA, and G blocks correspond to the schematic diagram of Fig. 2.7.

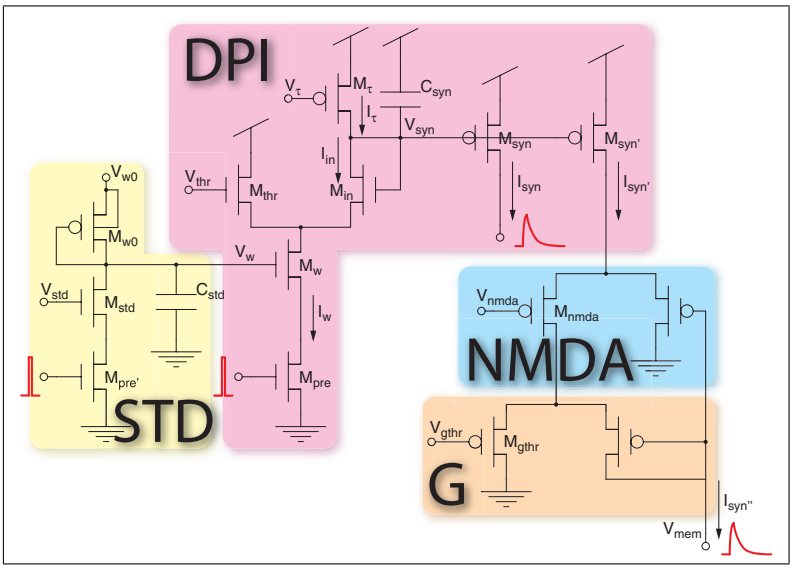


Figure 2.7: Schematic diagram of the DPI connected to additional circuits that augment the synapse's functionality. The names of the functional blocks correspond to the ones used in the layout of Fig. 2.6: The *STD* block comprises the circuit modeling short–term depression of the synaptic weight, the *NMDA* block comprises the transistors modeling NMDA voltage–gated channels, and the *G* block includes transistors that render the synapse conductance–based.

vary significantly, as it depends on the layout design. More conservative solutions use large transistors to obtain lower mismatch, but require more area. More aggressive solutions require less area, but multiple instances of the same layout cell produce currents with larger deviations.

Fig. 2.7 shows the schematic diagram of the full synaptic circuit implemented on the chip: the analysis of the previous section shows how the DPI response models the EPSC generated by biological excitatory synapses of AMPA type receptors (Destexhe et al., 1998); additional circuits can be attached to the DPI synapse to extend the model with additional features typical of biological synapses, and to implement various types of plasticity.

For example by adding two extra transistors, *voltage–gated channels* can be implemented to model NMDA receptor behavior. Similarly, by using two more transistors the synaptic model can become *conductance–based* (Kandel et al., 2000). Inhibitory ($GABA_a$) type receptors can be easily emulated by using the complementary version of the DPI circuit of Fig. 2.5, with a p–type diff–pair and n–type output transistor, as shown in Fig. 4.8.

The DPI circuit is also compatible with previously proposed circuits for implementing synaptic plasticity, both on short time scales with models of short–term depression (STD) (Rasche and Hahnloser, 2001; Boegerhausen et al., 2003), and on longer time scales with spike–based learning mechanisms, such as spike–timing–dependent–plasticity (STDP) (Indiveri et al., 2006). Finally, the DPI's extra degree of freedom for modifying the overall gain of the synapse either with V_{thr} or with V_w allows the implementation of synaptic homeostatic mechanisms (Bartolozzi and Indiveri, 2006), such as global activity dependent synaptic scaling (Turrigiano et al., 1998).

In the next paragraphs I first characterize the response of the DPI while disabling all the additional circuits, to validate the theoretical analysis of Sec. 2.2.6 by measuring the output current of the circuit when stimulated with variable voltage pulses, and with trains of pulses at different frequencies. I then describe the behavior of the additional circuits, characterized by measuring the membrane potential, V_{mem}, of a low power leaky integrate–and–fire (I&F) neuron (Indiveri et al., 2006) which receives as input the synaptic EPSC.

2.3.1 DPI response properties

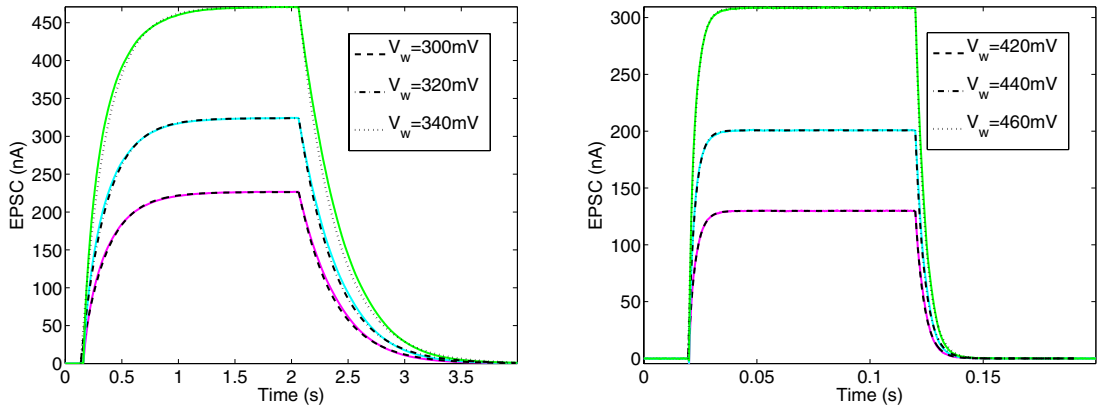
The DPI step response was measured for different synaptic weight, V_w, gain, V_{thr}, and time constant, V_τ. The synapse was stimulated with voltage pulses of varying duration, generated by a function generator. The currents measured from the DPI were transformed to voltages by means of an external current to voltage converter circuit. For each bias setting I repeated the same stimulation ten times and computed the mean and standard deviation of the response.

Fig. 2.8 shows the response of the synapse to an input pulse for different synaptic weight, V_w, bias values. The rise and decay parts of the data were fitted with the charge phase and discharge phase parts of eq. (2.23), for $t_i^- = 0$, and $I_{syn}^- = 0$, i.e.

$$I_{syn}(t) = \begin{cases} \alpha \left(1 - e^{-\frac{t}{\tau}}\right) & \text{(charge phase)} \\ \alpha e^{-\frac{t}{\tau}} & \text{(discharge phase).} \end{cases} \tag{2.27}$$

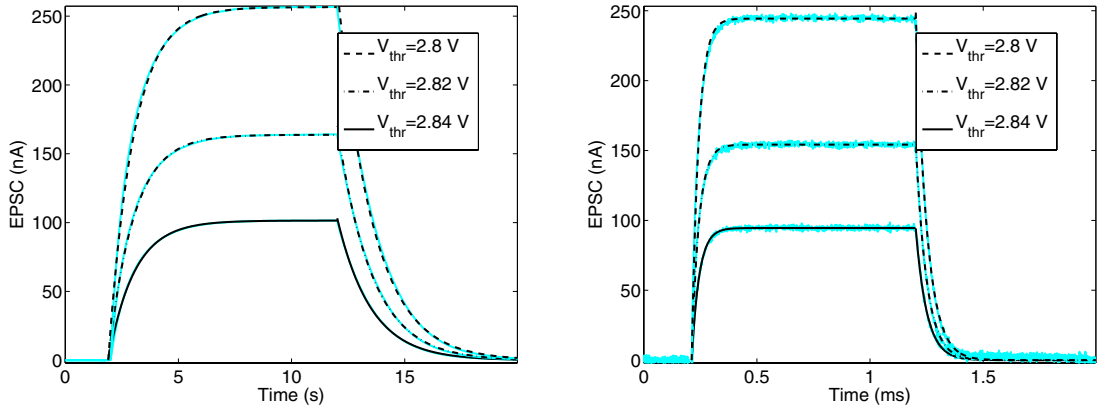
The small differences in the estimated time constants for the charge and discharge phases are most likely due to leakage currents and parasitic capacitance effects, not considered in the analytical derivations. These results however show that the DPI time constant does not depend on V_w, and can be independently tuned with V_τ. In particular, the time constants of the responses shown in Fig. 2.8(a) and 2.8(b) are compatible with the typical time evolution of biological NMDA and AMPA type receptors, respectively.

I also measured the DPI circuit's response to an input pulse for different synaptic gain, V_{thr}, bias values. Fig. 2.9(a) shows the circuit's response with a small I_τ current, as a function of different V_{thr} gain settings. The background shaded lines represent the measured data, while the solid, dashed and dot–dashed curves represent fits with eq. (2.27). Table 2.1 shows the fitting parameters, τ and α. The time constant estimated from the fits does not change with V_{thr} and is of the order of seconds. Fig. 2.9(b) shows the DPI response to input



(a) (b)

Figure 2.8: The response of the DPI to an input voltage pulse for two different values of V_τ and three different values of V_w. The response is fitted with eq. (2.23), and the fitting functions (dotted, and dashed lines) are superimposed to the measured data (shaded lines). (a) Slow time constant setting ($V_\tau = 3.1$V): the time constants estimated by the fit are $\tau = 219.3$ms for the charge phase, and $\tau = 326.3$ms for the discharge phase. (b) Fast time constant setting ($V_\tau = 2.94$V): the time constants estimated by the fit are $\tau = 2.9$ms, and $\tau = 4.1$ms.



(a) (b)

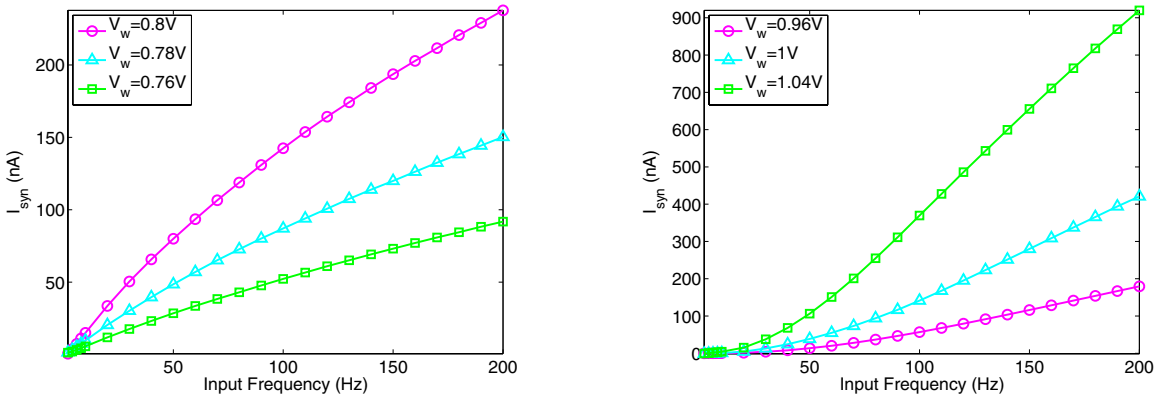
Figure 2.9: The response of the DPI to an input voltage pulse for two different values of V_τ and three different values of V_{thr}. The response is fitted with the eq. (2.23), and the fitting functions (dotted, and dashed lines) are superimposed to the measured data (shaded lines). In (a) I_τ is set to be very small (the p–FET used to generate I_τ has a $V_{gs} = 150$mV), and the circuit time constant is about one second. In (b) I_τ is set to be relatively large ($V_{gs} = 570$mV) in order to obtain a time constant of the order of μs (note the different time scale on the abscissa axis).

current pulses for larger values of I_τ, which produce time constants of the order of micro–seconds (note the different scale on the abscissa axis). These results are in accordance with both theoretical derivation and simulation results: decreasing V_{thr} increases the DPI gain exponentially, while the DPI time constant, set by adjusting the current I_τ, does not change with V_{thr}.

I also verified the derivation of the response of the DPI to spike trains in Sec. 2.2.6, by measuring the mean EPSC of the circuit in response to spike trains of increasing frequencies. Fig. 2.10 shows the $i - f$ curve for typical biological spiking frequencies, ranging

		$V_\tau = 2.77$V		$V_\tau = 3.15$V	
		Charge	Discharge	Charge	Discharge
$V_{thr} = 2.8$V	α(nA)	244.3934	248.6625	255.8926	271.1021
	τ(s)	30.0051μ	50.3472μ	1.0273	1.5578
$V_{thr} = 2.82$V	α(nA)	154.2117	155.7493	163.0349	170.1276
	τ(s)	31.4717μ	48.3105μ	0.9863	1.5178
$V_{thr} = 2.84$V	α(nA)	94.5382	94.3242	100.9033	103.1715
	τ(s)	32.8574μ	48.6947μ	1.0234	1.4874

Table 2.1: Fitting parameters of curves in Fig. 2.9. τ is the time constant and α the maximum of the curves, representing both the steady state of the exponential charge and the initial value of the exponential discharge.



(a) (b)

Figure 2.10: The response of the DPI to regular spike trains, for different synaptic weight biases V_w, and for the same two time constants of Fig. 2.8, (a) $V_\tau = 3.1$V, (b) V_τ=2.94V. The output mean current is approximately linear with the synaptic input frequency and its gain can be set with the synaptic weight bias V_w.

from 10Hz to 200Hz, for the two different values of the time constant bias V_τ used in the preceding experiments. The mean output current is approximately linear over a wide range of input frequencies (extending well beyond the ones shown in the plot). Fig. 2.10(a) shows that for long time constants the synapse response deviates from linearity due to the saturation effects described in Sec. 2.2.6. Fig. 2.10(b) shows a deviation from linearity for low input frequencies, when the time constant of the synapse is small; in such a case the synaptic weight is set to a high value, and the transistor M_w of Fig. 2.5 probably does not operate in the subthreshold regime any more, changing the regime of the DPI; nevertheless for high values of the input frequency the overall behavior is close to linearity.

2.3.2 NMDA functionality

The DPI reproduces the phenomenology of the current flow through ionic *ligand–gated* membrane channels, which open and let ions flow across the postsynaptic membrane as soon as they sense neurotransmitters released by the presynaptic boutons (e.g. AMPA channels). Another important class of *ligand–gated* synaptic channels, namely the NMDA receptors, are also *voltage–gated*; these channels open to let the ions flow only if the mem-

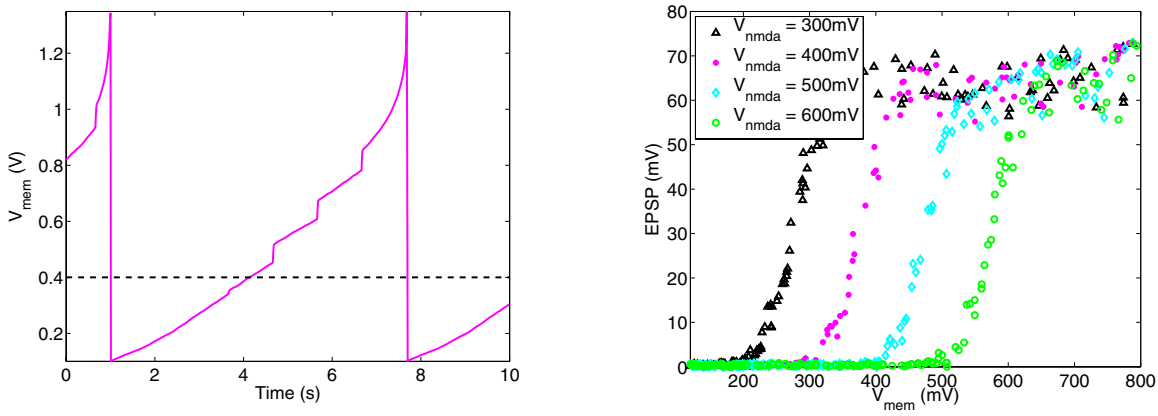


Figure 2.11: NMDA–type synapse response properties. (a) Membrane potential of an I&F neuron connected to the synapse and stimulated by a constant injection current. The NMDA threshold voltage is set to $V_{nmda} = 400$mV (dashed line in (a)). The small jumps in V_{mem} represent the excitatory postsynaptic potentials (EPSPs) produced by the synaptic input, when $V_{mem} > V_{nmda}$, in response to the presynaptic input spikes. (b) EPSP amplitude versus the membrane potential, for increasing values of the NMDA threshold V_{nmda}, and for a fixed value of V_w.

brane voltage is depolarized above a given threshold while in the presence of its neurotransmitter (glutamate). This voltage–gating behavior has been implemented by exploiting the thresholding property of the differential pair circuit (Mahowald and Douglas, 1991; Rasche and Douglas, 1999; Arthur and Boahen, 2004), as shown in Fig. 2.7: if the node V_{mem} is lower than the externally set bias, V_{nmda}, the output current, I_{syn}, flows through the transistor M_{nmda} in the left branch of the diff–pair, and has no effect on the postsynaptic depolarization. On the other hand, if V_{mem} is higher than V_{nmda}, the current flows also into the membrane potential node, depolarizing the I&F neuron, and thus implementing the voltage–gating typical of NMDA receptors.

Fig. 2.11 shows the results measured from the test circuit on the prototype chip; I stimulate the synapse with presynaptic spikes, while injecting constant current into the postsynaptic neuron's membrane. The amplitude of the synaptic EPSC depends on the difference between the membrane potential and the NMDA threshold, V_{nmda}. As expected, when V_{mem} is smaller than V_{nmda} the synaptic current is null and the membrane potential increases solely due to the constant injection current. As V_{mem} increases above V_{nmda}, the contribution of the synaptic current injected with each presynaptic spike becomes visible. The time–constant of the DPI circuit when used in this mode can easily be extended to hundreds of milliseconds (values typical of NMDA–type receptor dynamics) by increasing the V_τ bias voltage of Fig. 2.7. This permits the faithful reproduction of both the voltage–gated and temporal dynamic properties of real NMDA receptors. It is important to be able to implement these properties in VLSI devices because there is evidence that they play an important role in detecting coincidence between the presynaptic activity and postsynaptic depolarization for inducing long–term potentiation (LTP) (Morris et al., 1990). Furthermore the long time constant of the EPSC decay, easily tunable in the DPI implementation, is crucial for the function of the stabilizing role of the NMDA's receptor, which has been hypothesized by computational studies in the context of working memory (Wang, 1999) to be useful for stabilizing persistent activity of recurrent VLSI networks of spiking neurons.

2.3.3 Conductance–based functionality

So far, the total current flowing through the postsynaptic membrane channels has been shown to be independent of the postsynaptic membrane potential. However, in real synapses the current is proportional to the difference between the postsynaptic membrane voltage and the synaptic ion reversal potential E_{ion}, as given by

$$I_{syn} = g_{syn}(V_{mem} - E_{ion}). \tag{2.28}$$

Exploiting once more the properties of the differential pair circuit I can model this dependence with just two more transistors (see *G* block of Fig. 2.7), and obtain a behavior that, to a first order approximation, is equivalent to that described by eq. (2.28). Formally, the conductance–based synapse output is

$$I_{syn''} = I_{syn'} \frac{1}{1 + e^{\frac{\kappa}{U_T}(V_{mem} - V_{gthr})}}, \tag{2.29}$$

so by considering the first order term of the Taylor expansion, when $V_{mem} \cong V_{gthr}$ I obtain

$$I_{syn''} = \frac{I_{syn'}}{2} + g_{syn}(V_{mem} - V_{gthr}), \tag{2.30}$$

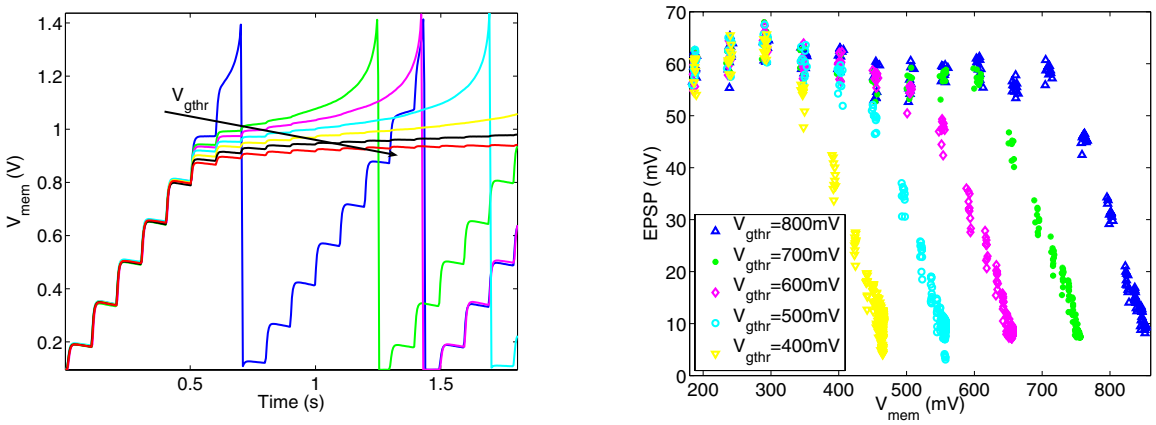
where the conductance term $g_{syn} = I_{syn'} \frac{\kappa}{4U_T}$.

Fig. 2.12 shows the EPSPs measured from the I&F neuron connected to the conductance–based synapse, for different values of V_{gthr}. These experimental results show that the synapse can reproduce the behavior of conductance–based synapses. This behavior is especially relevant in inhibitory synapses, where the dependence expressed in eq. (2.28) results in *shunting inhibition*. Computational and biological studies have attributed different roles to shunting inhibition, such as logical AND–NOT (Koch et al., 1983), or normalization (Carandini et al., 1997) functions. Evidence for these and other hypotheses continue to be the subject of investigation (Anderson et al., 2000; Chance et al., 2002). The implementation of shunting inhibition in large arrays of VLSI synapses and spiking neurons provides an additional means for exploring the computational role of this possible computational primitive.

2.4 Synaptic plasticity

In the previous sections I showed that the DPI circuit can model biologically realistic synaptic current dynamics. The main synaptic feature exploited in neural networks, though, is plasticity: the ability of changing the synaptic efficacy to learn and adapt to the environment. In neural networks with large arrays of synapses and neurons usually (Indiveri et al., 2006; Mitra et al., 2006; Arthur and Boahen, 2004; Shi and Horiuchi, 2004b) all the synapses belonging to one population share the same bias that sets their initial weight[3]. In addition each synapse can be connected to a local circuit for the short and/or long term modification of its weight. The DPI supports all of the short–term and long–term plasticity mechanisms for inducing short–term depression (STD), long–term potentiation (LTP), and long–term depression (LTD) in the synaptic weight that have been proposed in the literature. Specifically, the possibility of biasing M_w with subthreshold voltages in the order of hundreds of millivolts makes the DPI compatible with many of the spike–timing dependent plasticity

[3]The initial weight V_w can be set by an external voltage reference or by on–chip bias generators



(a) (b)

Figure 2.12: Conductance–based synapse response properties: (a) Membrane potential of the I&F neuron stimulated by the synapse, for different values of the synaptic reversal potential V_{gthr}.(b) EPSP amplitude as a function of V_{mem} for different values of V_{gthr}.

circuits previously proposed (Indiveri et al., 2006; Mitra et al., 2006; Arthur and Boahen, 2006; Bofill et al., 2002). The possibility of exploiting an extra degree of freedom for modifying the weight of the synapse allows the implementation of synaptic homeostatic mechanisms (Bartolozzi and Indiveri, 2006). In the following sections I describe the behavior of the synapse when connected to a short–term depressing circuit, then I analyze a possible implementation of the homeostatic plasticity.

2.4.1 Short–term depression

Short–term depression (STD) is a mechanism that modifies synaptic efficacy on a time scale of the order of hundreds of milliseconds to seconds. It is induced by the recent history of the presynaptic firing rate. Computational models have been proposed in the past to account for the experimental observations of modification of the synaptic efficacy and to investigate the computational role of such phenomenon (Abbott et al., 1997; Tsodyks and Markram, 1997; Chance et al., 1998; Chance and Abbott, 2001). The most evident property of STD is the enhanced response to transients and the adaptation to sustained stimulations; in contrast to postsynaptic forms of adaptations, STD is specific to a single presynaptic input: each neuron receives multiple afferent synapses, each one with its typical range of input frequencies, and each presynaptic terminal locally adapts its own efficacy, implementing a form of local gain control.

It has been shown that above a limit frequency, f_{lim}[4], the steady state synaptic efficacy is inversely proportional to the input firing rate (Tsodyks and Markram, 1997; Abbott et al., 1997). This implies that for input frequencies above f_{lim} the amplitude of EPSPs is inversely proportional to the input firing rate. Under the approximation that synapses add linearly, the average membrane depolarization is proportional to the product of the EPSP amplitude and the input firing rate, therefore for high frequencies it loses its dependence on the input firing rate (Abbott et al., 1997). This loss of sensitivity to sustained activity is complementary to the increase in sensitivity to variations of the firing rate: the amplitude

[4]The limit frequency f_{lim} is determined by the release probability of the synapse. In rat's cortical neurons f_{lim} has been estimated to be in an interval between ∼10Hz-25Hz (Tsodyks and Markram, 1997)

of the first EPSPs after a variation, Δf, of the input is proportional to the inverse of the initial input frequency, f, before depression further modifies synaptic efficacy; the transient response of the average membrane depolarization is proportional to the product of the EPSP amplitude and the input firing rate, Δf, resulting in a percentage variation of the input signal proportional to $\frac{\Delta f}{f}$. When a neuron receives input from two different synapses, one with low input firing rates, the other with high input rates, each one of the two synapses adapts differently, rendering the postsynaptic neuron sensitive to small variations in the input frequency for the synapse with low input range, and insensitive to the same absolute variation for the synapse with high input frequencies, reducing the responsiveness to random fluctuations of the input. This mechanism is also a form of logarithmic compression of the input for neurons that typically integrate the activity of thousands of synapses, each firing over a range of about 1-100Hz, keeping their firing rate within the same range (Abbott and Regehr, 2004). Consequences of synaptic depression are the removal of correlations in a single input train, to implement more efficient information coding (Abbott and Regehr, 2004), and detection of synchronous change of uncorrelated Poisson spike trains (Abbott et al., 1997). Chance et al. (1998) formalized a model of synaptic short–term dynamics, including fast and slow forms of depression. Slow forms of depression account for contrast adaptation in V1 neurons, while the fast form of depression accounts for rapid transient response, decrease of the response to periodic activity and non–linear temporal summation; their model produces a phase advance in the neuronal response which they propose as a candidate for the implementation of direction selectivity in V1 cells.

The DPI synapse is easily extended with the short–term depression circuit proposed by Rasche and Hahnloser (2001), where the synaptic weight decreases with increasing number of input spikes and recovers during periods of presynaptic inactivity. Quantitative considerations and comparisons to short–term depression computational models have been presented in Rasche and Hahnloser (2001) and Boegerhausen et al. (2003). Here I present a synthesis of the comparison of the circuit model with the computational model of Chance et al. (1998) presented in Boegerhausen et al. (2003).

Computational model

The synaptic efficacy is expressed by the term gD, where g is the maximum non–depressed efficacy, and $D\epsilon[0,1]$ is the depression value. After a presynaptic pulse, at time t_{sp}, D is updated in a multiplicative way by a constant depressing factor d, that is

$$D(t_{sp}^{+}) = dD(t_{sp}^{-}). \tag{2.31}$$

Then it recovers with a time constant τ_d, that is

$$\tau_d \frac{dD}{dt} = (1 - D). \tag{2.32}$$

At steady state, after presynaptic stimulation at rate f, the average depression is

$$\langle D \rangle = \frac{1}{1 + (1 - d)\tau_d f}. \tag{2.33}$$

The dependence of D on the inverse of the input firing rate is reflected to the steady state amplitude of the EPSPs, and on the average postsynaptic depolarization.

STD circuit

Boegerhausen et al. (2003) derived similar equations for the circuit in the STD block of Fig. 2.7. The current through the adaptive *Tobi element* (Delbrück, 1994), implemented by the diode transistor M_{w0} with bulk connected to the source terminal, is given by

$$I_{w0} = I_{0p} e^{\frac{k(V_{w0}-V_w)}{U_T}} \left(1 - e^{-\frac{(V_{w0}-V_w)}{U_T}}\right). \tag{2.34}$$

During the recovery phase transistors $M_{pre'}$ and M_{std} do not conduct, and the voltage V_w relaxes back to its initial (un–depressed) value V_{w0}, via the capacitor: $I_{w0} = C_{std}\frac{dV_w}{dt}$.

The current $I_{rf} \triangleq I_{0p} e^{\frac{k(V_{w0}-V_w)}{U_T}}$, and its differential $dV_w = -\frac{U_T}{k}\frac{dI_{rf}}{I_{rf}}$, substituted in eq. (2.4.1), lead to

$$\frac{C_{std}U_T}{kI_{0p}} d\left(\frac{I_{0p}}{I_{rf}}\right) = \left(1 - \left(\frac{I_{0p}}{I_{rf}}\right)^{\frac{1}{k}}\right) dt. \tag{2.35}$$

That becomes

$$\tau_p \frac{dD}{dt} = 1 - D^{\frac{1}{k}}, \tag{2.36}$$

where $\tau_p = \frac{C_{std}U_T}{kI_{0p}}$, and $D = \frac{I_{0p}}{I_{rf}} = e^{-\frac{k(V_{w0}-V_w)}{U_T}}$. The depression is null, i.e. $D = 1$, when the synaptic weight $V_w = V_{w0}$. The update rule for D results from a variation ΔV_w caused by the current through the transistor M_{std}, which turns on during a presynaptic pulse. The exponential dependence of D on V_w results in a multiplicative update rule for D, as in the model

$$D(t_{sp}^+) = dD(t_{sp}^-), \tag{2.37}$$

where d depends on the bias voltage V_{std}.

From the update and recovery rules one can derive the steady state value for depression; if $k = 1$ the circuit model is a faithful implementation of the computational model. However k is a process parameter that depends on the operating conditions of the MOS transistors; typical values range between 0.6 and 0.8. The non–linear eq. (2.4.1) can be simplified for $D \ll 1$, when the synapse is fully depressed; in such a case the equation becomes linear and

$$\langle D \rangle = \frac{1}{\tau_d (1-d) f}. \tag{2.38}$$

In the DPI the amplitude of the EPSC is proportional to I_w (see eq. (2.23)), which in turn is proportional to D. Therefore the steady state EPSP amplitude is inversely proportional to f.

Fig. 2.13 shows the EPSPs of the I&F neuron connected to the synapse, having activated the STD block of Fig. 2.7, for two different settings of the leak current on the I&F circuit. These results confirm the compatibility between the DPI and the STD circuits, and show qualitatively the effect of short–term depression. More quantitative experiments on the effect of STD are described in Sec. 4.3.5.

2.4.2 Synaptic homeostasis

The type of synaptic plasticity described in the previous paragraph locally modifies the synaptic weight of single synapses on a relatively short time scale. Other forms of plasticity act on longer time scales, and render neurons capable of learning patterns, associations, and

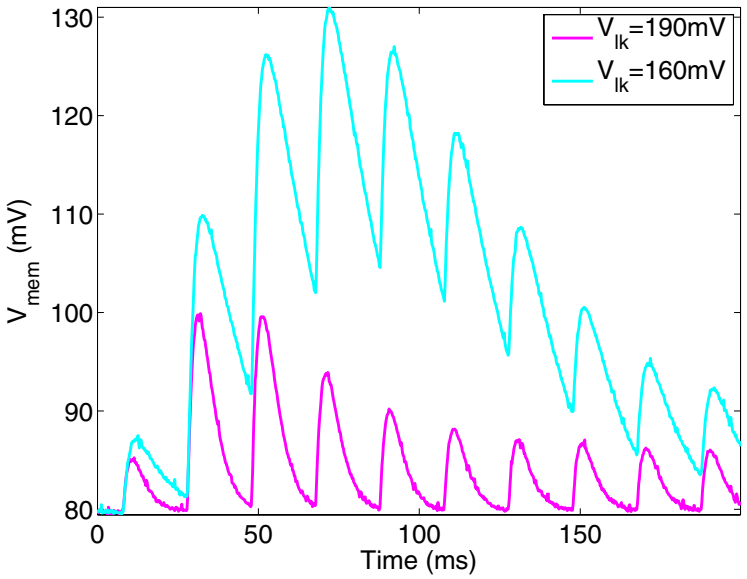


Figure 2.13: Short–term depression. Membrane potential of the leaky I&F neuron, when the short–term depressing synapse is stimulated with a regular spike train at 50Hz. The different traces of the membrane potential correspond to different values of the leakage current of the neuron. Note how (from the second spike on) the EPSP amplitude decreases with each input spike.

relations present in the external world. In this section I describe an additional type of synaptic plasticity that acts on even longer time scales, and is not synapse specific but collectively modulates populations of synapses. This type of plasticity mechanism is commonly referred to as homeostatic plasticity. It maintains the activity of neurons within a functional range, it stabilizes the activity of many interconnected neurons, and it reduces the effects of system inhomogeneities.

Physical implementations of systems comprising many similar elements operating in parallel, such as the brain or VLSI neural networks, have to cope with the intrinsic inhomogeneities present among different instantiations of the same elements. Specifically in VLSI, variations across the silicon wafer and small differences in doping concentration create differences between different instances of the same circuit. Similarly, in populations of real neurons, the different morphologies of each cell and their different distributions of ion channels and receptors lead to variations in their functional properties.

Despite their differences, real neurons constitute a homogeneous computational substrate, maintaining their overall level of activity within functional boundaries. Neurons show stable activity in the face of continuous turnover of their constituents, while retaining their ability to learn and adapt to new stimuli and changes in the environment. More importantly, real neurons can maintain stable activity also when they are part of complex highly interconnected networks. These stability and homogeneity properties are the result of various forms of homeostatic mechanisms that have been revealed in neurophysiology (Rutherford et al., 1998; Turrigiano, 1999; Desai et al., 2002; Burrone and Murthy, 2003). The specific mechanism I address here is referred to as *activity–dependent scaling of synaptic weights*: Turrigiano et al. (1998) showed how the level of activity in a neural population could be restored to its homeostatic value after inducing a chronic change, thanks to an "automatic gain control mechanism" that modifies the overall drive of the synapses in the network. Fig. 2.14 shows the effect of chronic increase or decrease of neuronal activity on the mean amplitude of EPSCs: the unveiled process acts by globally scaling the weights of

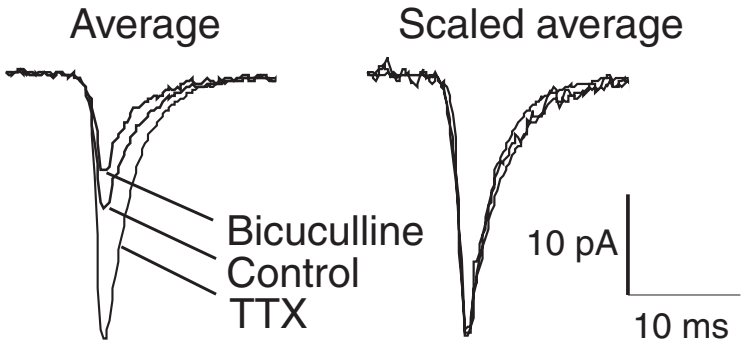


Figure 2.14: Activity–dependent scaling of synaptic weights: Mean EPSCs measured when the activity of a culture of neurons is chronically changed. Left: Tetrodoxine (TTX) causes decrease of spiking activity, and consequent increase in the amplitude of the EPSCs; Bicuculline has the opposite effect on spiking activity, that causes the decrease of the synaptic drive to the neurons. Right: Scaled EPSCs perfectly superimpose, demonstrating that only the amplitude, and not the dynamics, of the EPSCs is affected. Adapted from Turrigiano et al. (1998)

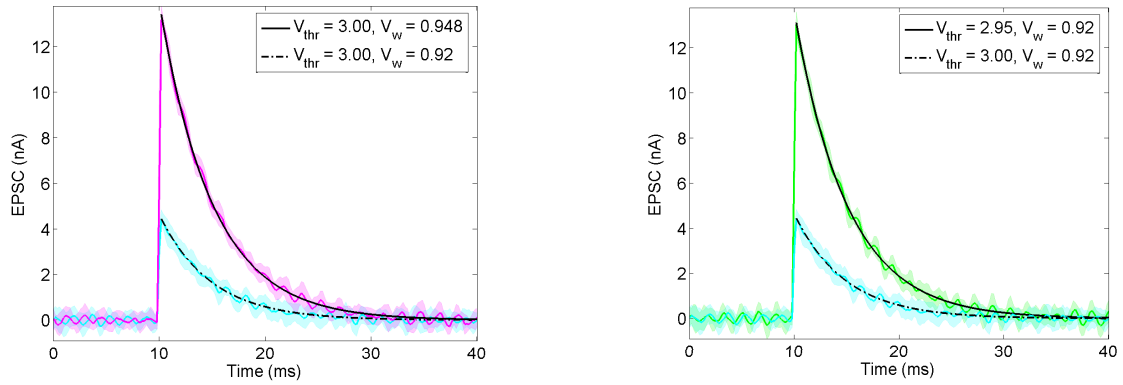
the entire distribution of inputs synapsing onto a single postsynaptic neuron, in response to chronic alteration of its output firing activity. The multiplicative nature of this mechanism preserves the relative differences between synaptic weights typically induced by local spike–based learning mechanisms. This type of homeostatic plasticity has been shown to exist both in neuronal cultures and *in vivo* during development (Desai et al., 2002); it regulates both the activity of single neurons and of entire networks (Rutherford et al., 1998). In the latter case it acts differentially on excitatory–to–excitatory synapses and on excitatory–to–inhibitory synapses, achieving stability by balancing the level of excitation and inhibition in the network.

The specific design of the DPI synapse allows the application of homeostatic control to VLSI implementations of neural networks, in order to compensate for inhomogeneities due to device mismatch and slow changes in the physical properties of the circuits arising due to temperature drift. In addition, in large multi–chip VLSI implementations of neural systems (Serrano-Gotarredona et al., 2005), instabilities could also arise also due to dramatic changes in the statistics of the input signals, induced for example by the incorporation of new input devices, by failures in existing sensory input devices, or by abrupt changes in the testing environment. In these situations the implementation of silicon homeostatic mechanisms could lead to improvements in overall system performance and stability.

For example in large aVLSI networks of spiking neurons (Indiveri et al., 2006), synaptic scaling can act on all afferent synapses of each neuron to maintain their activities within a functional range: this will naturally compensate for inhomogeneities across neurons caused by device mismatch. At the network level, homeostasis counteracts the effect of temperature drifts that can change the spiking activity of the neurons; at the system level it acts as an automatic gain control which responds to dramatic changes in input activity levels, i.e. when a chip is interfaced to a new sensory input device.

As demonstrated in Sec. 2.2.6, and shown in Fig. 2.15, the total synaptic efficacy of the DPI can be scaled by independently varying either I_w or I_{gain}. These two independent degrees of freedom can be exploited for learning the synaptic weight V_w with "fast" spike–based learning rules, while adapting the bias V_{thr} to implement homeostatic synaptic scaling on much slower time scales.

To test the feasibility of the homeostatic scaling with the DPI, I implemented homeostasis as a software control system, in loop with a chip comprising a VLSI implementation



(a) (b)

Figure 2.15: Independent scaling of EPSC amplitude by adjusting either V_{thr} or V_w. The plots show the time course of mean and standard deviation (over 10 repetitions of the same experiment) of the current I_{syn}, in response to a single input voltage pulse. In both plots the lower EPSC traces share the same set of V_{thr} and V_w, in (a) the higher EPSC is obtained by increasing V_w while in (b) by decreasing V_{thr}, with respect to the initial bias set. Superimposed to the experimental data are plotted theoretical fits of the decay from eq. (2.23). The time constant of all plots is the same and equal to 5ms.

of the synapse. In the following paragraph I describe the algorithm and show experimental data from the mixed–mode software/hardware neural system. The results obtained are favorable for the design of a hardware implementation of the algorithm.

Experimental setup and homeostatic control algorithm

This experiment involves the use of the DPI together with the I&F adaptive neuron, as previous experiments, but on a different chip and setup[5].

A Linux desktop was used to monitor the spiking activity of the I&F neuron in real–time, and to send sequences of spikes to the synapse (Oster et al., 2005; Dante et al., 2005). The desktop was also used to control a current source which injects a current I_n into the input capacitance of the I&F neuron, and to control a voltage source which sets the value of the DPI's V_{thr} bias voltage (see Fig. 2.5).

The neuron was stimulated using both current injection (sourced into the neuron's capacitance) and spike trains (sent to the DPI). The current I_n models the average input current that the neuron would receive from its full dendritic tree, and is used to induce a base activity level. The sequences of spikes represent the input to a single synapse, and could be used to drive a spike–based learning circuit such as the ones proposed by Indiveri et al. (2006) and Mitra et al. (2006).

To characterize the synaptic homeostasis model I fixed the statistics of the synapse input spike trains and varied the input current to the neuron, I_n. The homeostatic control algorithm adapted the DPI's V_{thr} bias to maintain the output firing rate of the neuron within a desired (functional) range. Formally, the control strategy adopted is that of a PI–controller: the algorithm determines how to change V_{thr} both by measuring the error between the firing

[5]The experiment described in this paragraph was performed on a different chip than the others presented in the thesis, implemented on standard 0.5μm technology and fabricated through the MOSIS consortium. The chip power supply is $V_{dd} = 5$V, instead of 3.3V

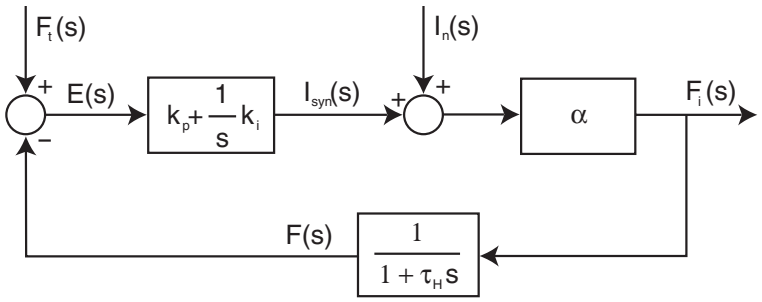


Figure 2.16: Block diagram of the homeostatic PI control algorithm, in the Laplace domain. $I_n(s)$, the disturbance input, and $I_{syn}(s)$, the system's controlled variable, are the current inputs to the I&F neuron. The feedback block integrates the neuron's output frequency $F_i(s)$ over time, the resulting low–pass filtered frequency $F(s)$ is then compared to the target frequency $F_t(s)$, generating the error $E(s)$ that drives the PI-controller block. It sets the controlled signal I_{syn} to a value that brings the neuron's output firing rate back to the reference value $F_t(s)$.

rate of the neuron and its target firing rate, and by computing the integral of the error over time. The block diagram of this classic control system is shown in Fig. 2.16.

The system of differential equations that implements this control strategy is

$$\begin{cases} \tau_H \dot{f}(t) = -f(t) + f_i \\ f_i = \alpha(I_n(t) + I_{syn}(t)) \\ e(t) = (f_t - f(t)) \\ I_{syn}(t) = k_p e(t) + k_i \int_0^t e(\xi) d\xi \end{cases}, \tag{2.39}$$

where τ_H is the time constant of the homeostatic process, f_i is the measured instantaneous firing rate of the neuron, α is the transfer function gain of the neuron, when operating in its linear region (Indiveri, 2003b), f is the integrated neuron firing rate, and f_t is a desired target firing rate.

This control algorithm determines the value of I_{syn} required to keep the output firing rate close to a defined target rate; the updated value of I_{syn} depends proportionally on the the error of the firing rate, $e(t)$, and on its integral over time, with the proportionality constants k_p and k_i respectively. To set I_{syn} to the new desired value, I use eq. (2.26) and modify I_{gain} (via V_{thr}) accordingly.

This software algorithm can be directly mapped on silicon: another instance of the DPI circuit could be used to implement the integration of the output firing rate over time, a differential pair can be used to realize the proportional control, and a follower integrator circuit can be used to implement the integral control.

Experimental Results

To demonstrate the properties of the homeostatic control setup I replicated the experiment described by Turrigiano et al. (1998), where they chronically shifted the activity of a population of neurons to induce synaptic scaling behavior. Specifically, I initially combined current injection and synaptic stimulation such that the neuron fired at a desired rate of approximately 98Hz. Subsequently I produced a step change in the I&F neuron's firing rate by changing the injection current I_n, and let the control algorithm scale the total synaptic efficacy. As shown in Fig. 2.17, the homeostatic control adapted the neuron's firing rate

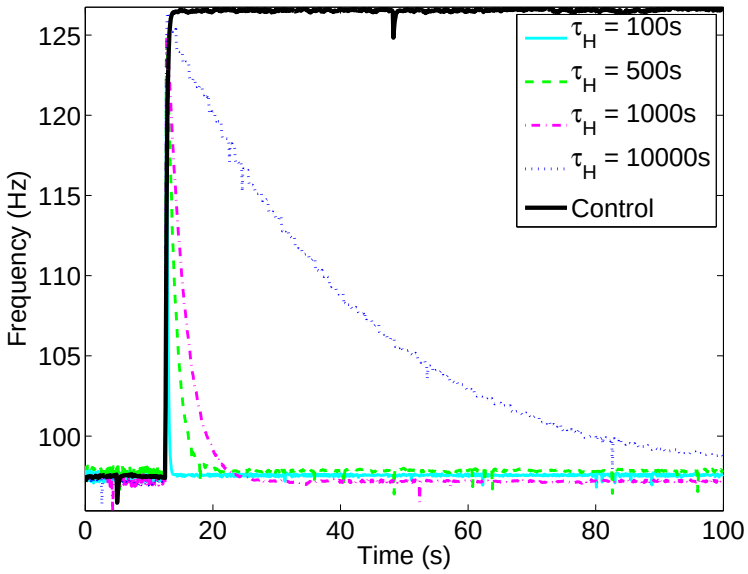


Figure 2.17: Homeostatic response to a step–wise DC shift in the neuron's instantaneous firing rate. The thick black line shows the output of the neuron for a step in the input current level when the homeostatic control is not enabled. The other curves show how the firing rate goes back to the initial activity level for different time constant of the homeostatic control.

back to its target value with a time constant τ_H. In the experiment shown in Fig.2.17, the control algorithm adapted the V_{thr} bias from a value of 4.5V to one of 4.58V. This produced a decrease of I_{gain}, which in turn scaled the amplitude of EPSCs proportionally, reproducing the behavior observed in (Turrigiano et al., 1998). The homeostatic control algorithm is symmetrical: decreasing the current injection level results in a decrease of activity that recovers to the initial value with the same time constant as the previous experiment. In this case the change in activity has been achieved by decreasing V_{thr}, thereby increasing the amplitude of EPSCs with respect to their control values.

Ideally, the (slow) homeostatic stabilizing mechanism should not interfere with the (fast) spike–based learning mechanisms. To show that the homeostatic control algorithm corrects only chronic DC and low frequency shifts of activity, allowing the information associated with fast fluctuations of the input signal pass through, I superimposed high–frequency fluctuations on I_n and repeated the chronic (step) change experiment. Fig. 2.18 shows the results of this experiment. As shown, the DC offset is removed while the high frequency fluctuations are transmitted by the I&F neuron. The amplification of the high–frequency components is due to the choice of the k_i, and k_p parameters in the control algorithm.

2.5 Applications to biomedical signal processing

In the previous section of this chapter I described how it is possible to apply principles observed from biology to solve technological issues. In this case the mechanism of homeostatic plasticity was implemented by applying a solution from classical control theory.

In this section I demonstrate that the design of circuits like the DPI, carried out in the context of neuromorphic analog VLSI research with the aim of realizing a faithful VLSI models of biological synapses, can lead to the design of novel circuits (in this case a log–

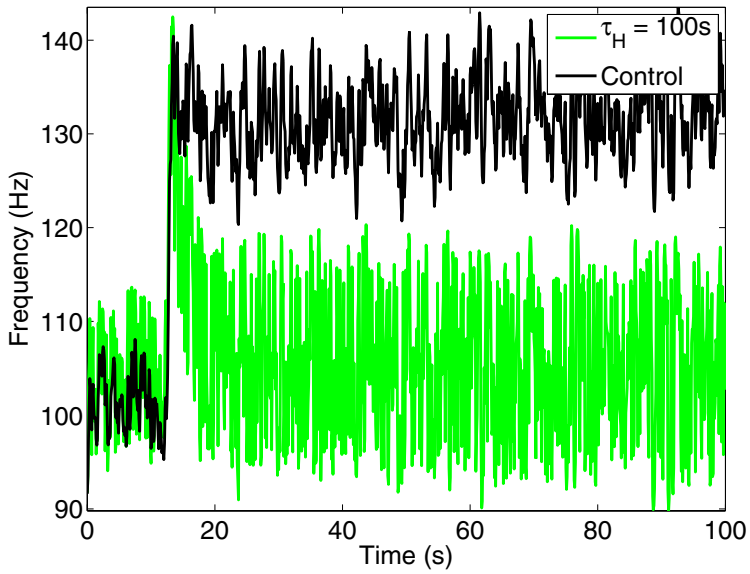


Figure 2.18: Homeostatic control adding high frequency fluctuations to the injection current. I replicate the same experiment of Fig. 2.17, adding random noise, for the time constant $\tau_H = 1000$s; the black line shows the output of the neuron for a step in the input current level when the homeostatic control is not enabled. The blue curve shows how the DC offset in the firing rate is corrected, without affecting high frequency fluctuations in firing rate.

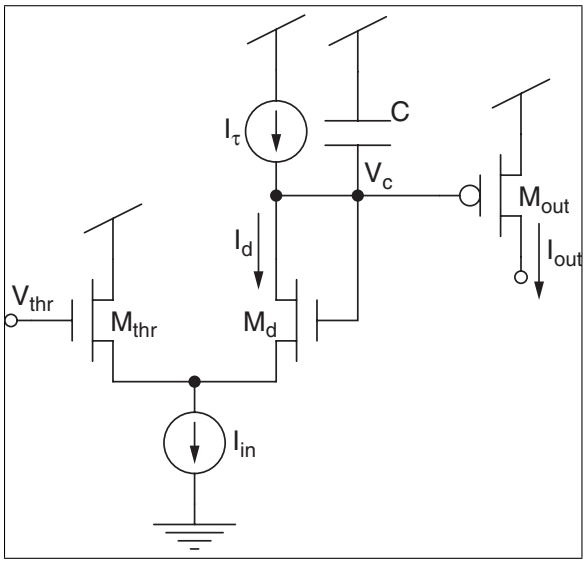


Figure 2.19: Current–mode log–domain DPI integrator.

domain low–pass filter) suitable also for other more traditional engineering application domains such as biomedical signal processing.

When the input to the DPI circuit is the current $I_{in}(t)$, as shown in Fig. 2.19, the DPI behaves as a linear integrator with the same properties of classical log–domain first order low pass filters (Frey, 2000), but with the additional advantage of providing tunable gain independent from the (tunable) time constant, a compact layout, better matching properties and lower power consumption.

As already mentioned, current–mode circuits have been shown to have a wide variety

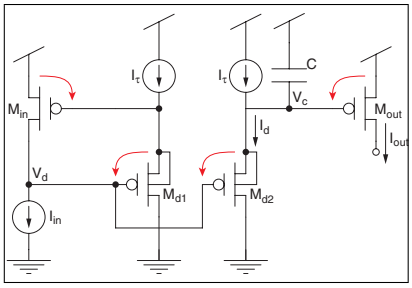


Figure 2.20: Classical log–domain first order linear filter. The arrows show the V_{gs} loop used for the translinear principle application.

of useful features, including the capability of operating with large bandwidth at low supply voltages (Ramirez-Angulo et al., 1992). Current–mode CMOS circuits operated in the subthreshold, or weak–inversion, regime can be used to implement *log–domain filters* (Frey, 2000). The log–domain paradigm has the advantage of producing linear building blocks by dealing with non linearities at the component level and, as any other companding technique (Seevinck, 1990), it improves the circuit's dynamic range (Frey, 1993).

In the next sections I describe the DPI properties, when analyzed from the point of view of low–pass filtering, and compare it to a classical example of a low–pass filter proposed by Frey (2000). I report the results of SPICE simulations characterizing the circuit transfer function and parameters evaluating the filtering performance, such as the Total Harmonic Distortion and power dissipation.

2.5.1 Low–pass filtering

In Sec. 2.2.6 the theoretical circuit analysis lead to the derivation of the circuit transfer function (eq. (2.22)); a first–order *non–linear* differential equation.

If the DC component of the input signal, I_{in}, is much greater than I_τ, then $I_{out} \gg I_{gain}$. Under this condition, the second term of eq. (2.22) reduces to $I_{in}\frac{I_{gain}}{I_\tau}$ and I obtain a *linear* first order differential equation characteristic of linear filters, but with tunable gain $\frac{I_{gain}}{I_\tau}$. In the Laplace domain the DPI transfer function is therefore:

$$\frac{I_{out}}{I_{in}} = \frac{I_{gain}}{I_\tau} \cdot \frac{1}{1+\tau s} \tag{2.40}$$

Fig. 2.9 shows the independent control of the time constant and gain of the circuit, and the fits of the curves confirm the linearity of the system.

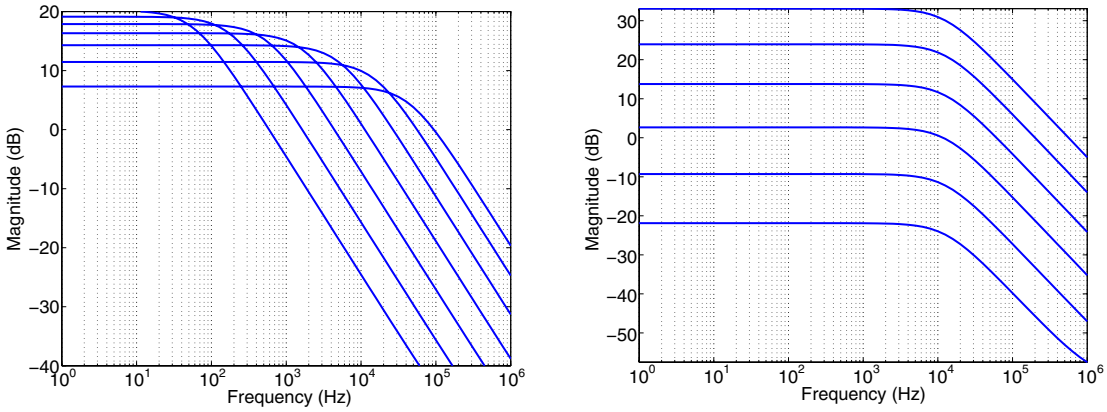
For comparison, Fig. 2.20 shows the classical log–domain integrator proposed by Frey (2000), with equivalent functionality. This circuit's linear transfer function can be easily derived by applying the translinear principle on the V_{gs} loop highlighted by the arrows in Fig. 2.20: given the exponential relationship between the subthreshold currents of the p–FETs and their V_{gs} voltages, I can write: $I_{in} \cdot I_\tau = I_{out} \cdot I_d$. Similar to the DPI analysis, differentiating I_{out} with respect to V_c and combining the result with the capacitor equation $C\frac{d}{dt}V_c = -(I_d - I_\tau)$, I derive the standard first order differential equation

$$\tau\frac{d}{dt}I_{out} + I_{out} = I_{in}. \tag{2.41}$$

As this circuit requires p–FETs with isolated wells it occupies more silicon area than the DPI.

Table 2.2: Dimensions of elements used in simulations and in the circuit implementation. The MOSFET entries show their W/L values expressed in μm/μ), while the (MOSCAP) capacitor area yields a capacitance of 770fF.

M_{in}	6.3/3	M_g	1/3
M_d	1/3	M_τ	1.7/2.2
M_{out}	13.5/2.2	C	170μm^2



(a) (b)

Figure 2.21: The simulated DPI circuit transfer function, for a DC input current value of 10nA; (a) plots with values of I_τ ranging from 0.3nA to 1.5nA; (b) plots with values of V_g ranging from 2.6V to 2.85V.

SPICE simulations of the DPI circuit were performed[6] using 0.35μ m AMS process parameters, with both 3.3V and 1.2V power supply settings. The transistor dimensions and capacitance value used for the simulations match those in the layout of Fig. 2.6 and are listed in Tab. 2.2.

Fig. 2.21 shows the AC simulation results, characterizing the properties of the DPI in the frequency domain as a function of its time–constant and gain. Fig. 2.21(a) shows the simulation results for different values of the DPI time constant, obtained by changing the current bias I_τ. Even with relatively small capacitor values (see Tab. 2.2), the DPI integrator can produce time constants of values as long as hundreds of milliseconds, providing a very low cut–off frequency low–pass filter. The gain of the DPI transfer function decreases in a geometric fashion when the cut–off frequency increases linearly, because I_τ appears in the denominator of the transfer function (see eq. (2.40)). Fig. 2.21(b) shows the simulation results for different values of the bias voltage V_{thr}, which is modulated to change the gain of the DPI. The simulations confirm the theoretical analysis: different settings of V_{thr} affect the circuit's gain, while leaving the cut–off frequency unchanged. Analogous results have been obtained when measuring the step response of the DPI circuit from the fabricated prototype chip (see Sec. 2.3.1).

To test the linearity condition derived in Section 2.2.6, the DPI circuit was stimulated with input currents with a DC component I_{in} greater than I_τ, and different values of AC component i_{in}. In the simulations $V_{thr} = V_{dd}-$ 0.4V, $I_\tau =$ 1pA, $I_{in} =$ 10pA, and the frequency of the AC input signal was matched to the filter cut–off frequency of 6Hz. Fig. 2.22

[6]The simulations were kindly performed by S. Mitra and G. Indiveri

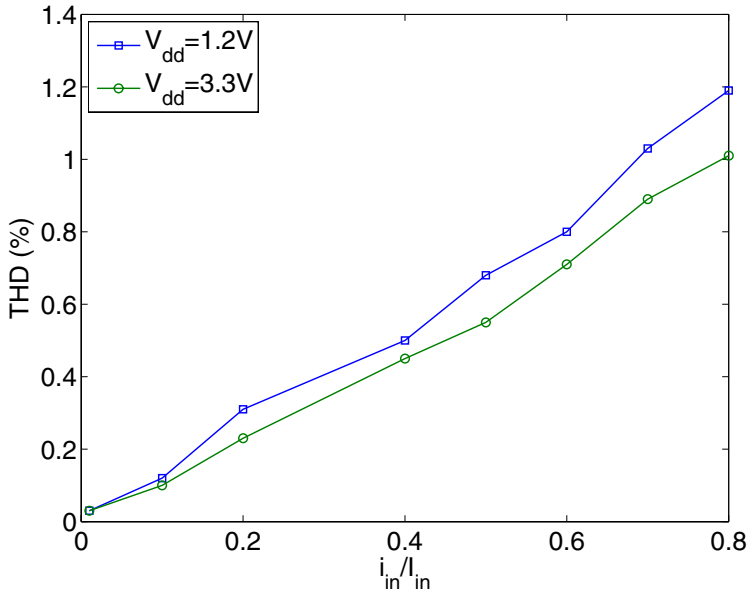


Figure 2.22: Simulated Total Harmonic Distortion (THD) of DPI circuit, for two values of the supply voltage V_{dd}.

Table 2.3: DPI circuit specifications.

Area (without pads and guard rings)	464.750μm^2
Power dissipation @ $I_\tau = 1$pA, $V_{dd} = 1.2$V	0.7nW
Supply voltage	1.2V-3.3V
f_c tuning range	from 1Hz - 50KHz
THD @ i_{in}/I_{in} = 0.1	-60dB
THD @ i_{in}/I_{in} = 0.8	-41.4dB

shows the circuit's Total Harmonic Distortion (THD) as a function of i_{in}/I_{in}, for two different values of supply voltage.

The values used in the simulations above are typical in neuromorphic and biomedical applications (Indiveri et al., 2006; Sol'is-Bustos et al., 2000). In these conditions (and with $V_{dd} = 1.2$V) the circuit dissipates less than 1nW. Fig.2.23 reports the power dissipation of the DPI as a function of desired cut–off frequency. In this experiment $V_{thr} = V_{dd}-0.4$V, the i_{in}/I_{in} ratio was fixed to 0.5 (for a THD of approximately 0.6%), I_{in} was set to $10I_\tau$, and I_τ was varied between 1pA and 50nA. The cut–off frequency was computed for each value of I_τ, stimulating the DPI with the same frequency and measuring the average power dissipation. As shown, the power consumption is proportional to the desired cut–off frequency (*i.e.* to I_τ), and for frequencies lower than 100Hz it is extremely efficient.

2.6 Conclusions

In this chapter I described the Diff–Pair Integrator, a new neuromorphic circuit that models one of the basic elements of neural wet–ware: the synapse. The postsynaptic currents produced by the DPI evolve exponentially with time and are a good approximation of real synaptic conductances, as demonstrated by the computational model proposed by Destexhe et al. (1998).

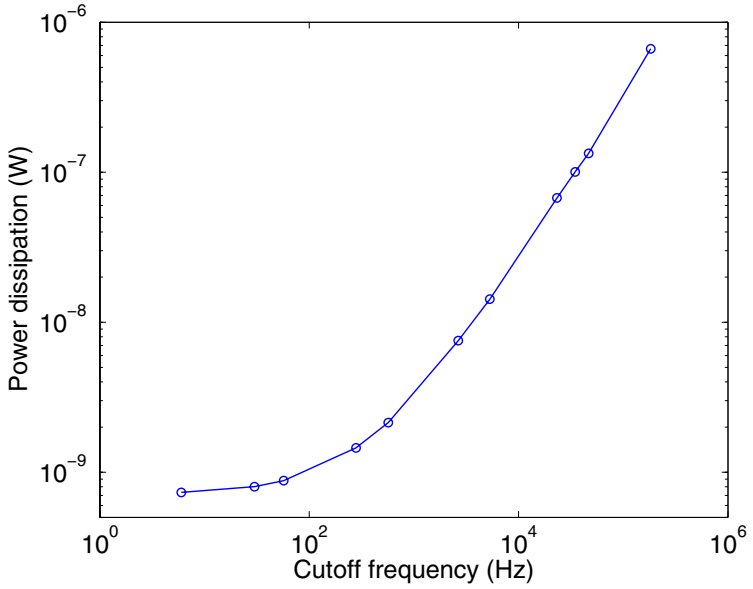


Figure 2.23: Simulated power dissipation for increasing values of cutoff–frequency (set by I_τ)

I compared the DPI with previous silicon synaptic emulations, highlighting the improvements introduced by the new design. At the circuit level the DPI is very compact and allows a better control over the synaptic parameters. The exponential characteristic of the circuit leads to a linear summation of the effects of input spikes over a wide range of input frequencies, being especially relevant for the implementation of dense arrays of synaptic matrices, where each DPI can be used to spatially sum the activity of many presynaptic neurons.

At the modeling and computational level the DPI offers many opportunities, of which I underlined the relevance and role in synaptic computation. For example, the possibility of easily tuning the time constant of the synaptic currents to a very slow decay has an important implication in maintaining persistent activity in recurrent networks at low firing rates, a mechanism correlated with the existence of working memory (Wang, 1999).

With the addition of a few extra transistors, the synaptic model can be extended to include voltage–gating typical of NMDA receptors and conductance–based current flow. The former is important as a coincidence detector of pre and postsynaptic activity, the latter is crucial for the design of inhibitory synapses and shunting inhibition.

I demonstrated that the DPI circuit is compatible with existing circuits for short and long–term modification of synaptic efficacy. The proposed synapse gives the possibility of implementing those mechanisms of adaptation and learning, such as short–term depression and spike timing dependent plasticity, that are the foundations of real neuronal computation, allowing the system to interact with its environment.

The mentioned extra degree of freedom for the synaptic weight is especially relevant for the implementation of an additional type of plasticity rarely modeled in silicon systems: synaptic homeostasis. It is an adaptive mechanism that globally scales synaptic weights over a very long time scale, to keep neuronal activity within functional boundaries. As a result it equalizes the output range of neurons, reducing their mismatch and adapting to chronic changes in the environment and stimulus range.

In summary, I presented a new synaptic circuit with additional modules that introduce extra functionalities to the basic model. I underlined the computational relevance of the

additional features and the importance of their inclusion in future chip design.

The proposed silicon synapse with all of its extensions enriches the vocabulary of neuromorphic VLSI computational primitives, with particular attention to different forms of adaptation. It gives the possibility of implementing within a single framework, on dense arrays of synaptic matrices, many aspects of synaptic computation that are likely to introduce crucial functionality into silicon emulations, moving a step ahead in the neuromorphic quest for the understanding and implementation of computational principles and strategies of the neural substrate.

In the last section of this chapter I analyzed the circuit from a classical engineering point of view, showing that it is a first order low pass filter with good performance in terms of power consumption for very low cut–off frequencies. I suggest that it can be of interest in the biomedical field, showing how neuromorphic research can lead to the design of efficient circuits that can be utilized in current technology.

The following chapters will illustrate the relevance of the DPI synapse to neurally–plausible computational principles, by describing its use in the implementation of a selective attention model.

Chapter 3

Silicon Winner–Take–All circuits

3.1 Introduction

In this chapter I describe the circuit that forms the computational core of the Selective Attention Chip, a bi–dimensional current–mode hysteretic Winner–Takes–All (WTA). It selects the pixel receiving the highest input current, while suppressing all the others.

In cortex there are circuits of neurons with recurrent connectivity that exhibit WTA properties (among others) (Douglas et al., 1994). WTA circuits are basic building blocks of artificial neural networks; Maass (2000) showed that soft WTA[1] networks can approximate any continuous function, and k–WTA[2] networks applied to the weighted sum of the input variables can compute any boolean function. Moreover, given the same implementation cost, WTA networks are more powerful than multi–layer perceptrons for computation; for example, the perceptron needs n^2 gates to perform WTA computation of n input variables, while WTA circuits need only n gates (Lazzaro et al., 1989). Hahnloser et al. (2000) have shown that soft WTA circuits perform simultaneously both digital selection and analog amplification. They show that for a given set of inputs the response of the circuit converges to a solution that selects the region of highest input activity, while its amplitude is graded and depends on the input strength and on the overall input activity level, exhibiting analogue properties such as gain modulation. This functionality can implement attentive modulation effects on visual responses (Lee et al., 1999): when a region of interest is selected, the neurons respond to the sensory characteristics of the input, and their activity is modulated by attention. The WTA circuit proposed by Hahnloser et al. (2000) also shows coexistence of analogue response and multi–stability: when one of many stimuli is selected by attention at the expenses of the others, the response to the stimulus is graded by the sensory characteristics of the selected stimulus, as it was presented alone (see Reynolds and Chelazzi (2004) for a review). When such a network is stimulated by two spatially separated stimuli it is capable of selecting one, while averaging between two proximal stimuli. In this latter case the circuit interprets the two close stimuli as a corrupted version of a single stimulus, and restores its output towards this interpretation. Chicca (2006) demonstrates that a recurrent circuit similar to the one proposed by Hahnloser et al. (2000), but implemented using spiking units, shows the same soft WTA computation, hysteresis, and sharpening of the input; properties typical of cortical computation (Douglas et al., 1999). I carried out software simulations of the network as proposed in (Chicca, 2006; Hahnloser et al., 2000), implemented with non–spiking Linear–Threshold Units (LTU). The results, described in Appendix A,

[1] soft WTA networks outputs analog numbers, whose values depend on the rank of the corresponding input.

[2] k–WTA networks compute a function that assigns the value 1 to the output corresponding to the k strongest inputs, and 0 to all others.

show that the network is capable of reproducing the behavior observed in the mentioned works; additionally those experiments assess the robustness of the WTA circuit and of its computational characteristics, by exploring its behavior when under changes in the relative strength of recurrent connectivity.

3.2 State–of–the–art WTA silicon implementations

The cooperative/competitive type of computation performed by recurrent WTA networks has inspired extensive research in the field of artificial neural networks. This approach represents a good alternative to classical engineering strategies, in solving tasks that require processing large amounts of fuzzy, noisy, real world data, such as pattern recognition and classification (Choi and Sheu, 1993). In particular, hardware implementations of WTA networks have been extensively applied to image compression (Choi and Sheu, 1993; Demosthenous et al., 1998) and pattern recognition problems, for example the design of Hamming networks (Robinson et al., 1992); it is especially useful for applications involving speech (Mead et al., 1991) and image processing (Mahowald, 1994; Häfliger and Bergh, 2002), and it is the circuit of choice for implementing systems for selective attention and tracking (Indiveri, 2003a; Morris et al., 1996; Horiuchi and Koch, 1999; Brajovic and Kanade, 1998). The WTA circuit is used where the hardware system must take a decision. For example in (Mead et al., 1991) it is used to determine if the auditory signal from the left and right ears are correlated or anti–correlated, by comparing the strength of the correlation and anti–correlation signal for each pixel. In (Mahowald, 1994) it is used to suppress the response to false correspondences in binocular vision, realizing an elegant solution to the problem of stereo–correspondence in binocular depth perception. The decision power of the WTA circuit has another important field of application in competitive learning networks, where only the unit winning the competition can modify its synaptic weight (Perfetti, 1990; Choi and Sheu, 1993).

For the above mentioned applications, and in particular for the modeling of selective attention and tracking, not all of the properties of the network proposed in (Hahnloser et al., 2000) are necessary; a hard WTA with one single output is sufficient for such tasks, since the unique output can be directly used to signal where attention is deployed.

One of the first analog VLSI circuits implementing the hard WTA function was proposed by Lazzaro et al. (1989); since then several different circuits and modification of the original circuit have been proposed, each optimizing different characteristics. The main constraints imposed by the hardware realization of arrays that can process a large number of inputs are compactness, power consumption, resolution, and speed; Table 3.1 summarizes the most popular WTA circuits, emphasizing these characteristics.

The proposed circuits can be divided into current–mode and voltage–mode circuits, and between asynchronous and clocked circuits.

Current–mode circuits have been shown to have a wide variety of useful features, including compactness, low power consumption, and wide dynamic range at low supply voltages (Ramirez-Angulo et al., 1992).

The choice of asynchronous circuits stems from our wish to model biological systems, which are intrinsically asynchronous and event driven, with computation performed continuously in time. This approach is compliant with other neuromorphic devices, such as the

	Mode	Input Output	Size	Resol.	Proc. time	Power cons.	Characteristics	Array
Lazzaro et al. (1989)	i (sub)	i $v \propto ln(i)$	3T $o(N)$	10% $\downarrow$ N	$\uparrow$ N	$\uparrow$ N	optimal compactness, suffers from mismatch when N increases	16×1
Indiveri (2001a); DeWeerth and Morris (1995)	i (sub)	i i and $v \propto ln(i)$	7T $o(N)$	1% $\downarrow$ N	$\uparrow$ N	$\uparrow$ N	increased resolution and speed, hysteresis, local excitation and inhibition (+2T), suffers from mismatch when N increases	25×1
Fish et al. (2005)	i (ab.) i (sub)	i v	$12T$ $o(N)$	30nA 1.8nA	8ns 34ns	87.5μW 22.5μW per cell	high speed and precision, adaptive threshold, hysteresis (too strong, needs reset), very high power consumption	8×1
Demosthenous et al. (1998)	i (above)	i v and $i \propto i_{max}$	$25T$ $o(N)$	10% $\downarrow log_2 N$	223ns (N=1024) $\uparrow log_2 N$	100μW per cell $\uparrow log_2 N$	high power consumption, large area, tree structure, high precision and speed, effect of mismatch doesn't increase with N	8×1
Choi and Sheu (1993)	v (above)	v v	10T×2 $o(N)$	15mV	$\sim$100ns $\uparrow$ N	>250μW per cell	high precision and speed, low degradation due to mismatch, multi-chip array, high power consumption	50×1 per chip
Serrano-Gotarredona and Linares-Barranco (1998)	i (above)	i v and i_{max}	6T 1C $o(N)$	2.35% to 0.56% – N	$\sim$10μs to $\sim$60ns $\uparrow$ N	$\uparrow$ N	high resolution, no degradation due to mismatch, multi-chip array, stability problems	10×1 per chip

Table 3.1: Characteristics of popular aVLSI WTA circuits. Only one of the reported circuit is voltage-mode (v), the others are current–mode (i) circuits, operated either in subthreshold (sub), or in strong inversion (above). The number of transistors (T) and capacitors (C) give an indication of the relative sizes. All of the circuit but one have $o(N)$ complexity, where N is the total number of inputs. Some circuits suffer from mismatch and their performance degrades with increasing number of inputs. The most precise circuits, less prone to degradation, are less compact and consume more power. The field "array" shows the number of inputs of the arrays fabricated and tested.

silicon retina (Lichtsteiner et al., 2006b) or cochlea (van Schaik and Liu, 2005), which can be naturally interfaced with the WTA–based chip described in this thesis via a common asynchronous protocol (Deiss et al., 1998; Chicca et al., 2006b) (see Sec. 4.2.1). Besides these motivations, the additional circuitry related to clocked systems increases power consumption.

Table 3.1 reports the main characteristics of some of the most popular asynchronous WTA circuits. The circuits listed are based on current–mode design, with the exception of the one proposed by Choi and Sheu (1993), which exploits voltage–mode design to reduce the impact of mismatch. In all other circuits listed, the input signal is a vector of currents; the output can be either voltage or current. In all of them the output is a binary encoding of the result of the competition: the value corresponding to the winner is high and all others are low. In some cases, a second output encodes the analog value of the winning input: in (Lazzaro et al., 1989; Indiveri, 2001a) the output current is either zero, or equal to a bias current when the corresponding cell wins, and the output voltage of the winning cell is proportional to the logarithm of the maximum input current; in (Demosthenous et al., 1998; Serrano-Gotarredona and Linares-Barranco, 1998) the output voltage is binary and the output current is proportional to the input current of the winner. The size (silicon area) of each circuit is not directly comparable, since the circuits are fabricated with different technologies; the number of transistors per cell gives a rough estimate of the relative size among different circuits, although this number cannot account for differences due to varying transistor size, a critical element in analog circuit design. Günay and Sánchez-Sinencio (1997) made a comparison among some of the circuits listed in Table 3.1 by fabricating them in the same technology: the circuit proposed in (Lazzaro et al., 1989) is the smallest, the one in (Choi and Sheu, 1993) is almost twice as big, and the one in (Serrano-Gotarredona and Linares-Barranco, 1998) is more than five times the first one, but twice as fast and precise. All the proposed circuits are fully parallel systems: the N input currents go to as many instances of the WTA circuit which usually compete via a single common node, implementing the global inhibition mechanism. This structure has $O(N)$ complexity: size, connectivity, and power consumption scale linearly with the number of inputs N. An exception to this scheme are the tree structures (Demosthenous et al., 1998; Wawryn and Strzeszewski, 2001), where the competition is performed by a hierarchy of 2–input WTA cells; in this case the system has $O(N)$ complexity, but size, power consumption and processing time grow proportionally with the logarithm of the number of inputs N.

One big concern in classical engineering applications is precision. The performance of circuits similar to (Lazzaro et al., 1989) in terms of resolution decreases with the number of inputs due to mismatch among transistors; increasing transistor size, the use of feedback (Indiveri, 2001a), special design techniques such as cascoding (Serrano-Gotarredona and Linares-Barranco, 1998), or the operation of the circuits in the strong inversion regime limit the effect of mismatch, at the expenses of silicon area required and/or higher power consumption. In general, processing time is faster when the input currents are high and the separation between the maximum input and the others is higher (Indiveri, 2001a; Fish et al., 2005; Demosthenous et al., 1998; Serrano-Gotarredona and Linares-Barranco, 1998).

The circuit proposed by Lazzaro et al. (1989), together with the modifications introduced by Starzyk and Fang (1993) and Indiveri (2001a), optimizes silicon area usage and power consumption. It is ideal for tasks that do not require high resolution or high speed, such as sensory perception tasks like the modeling of selective attention described in this work.

In the next section I describe the circuit implemented on the SAC, beginning from the original circuit proposed by Lazzaro et al. (1989) and describing the modifications intro-

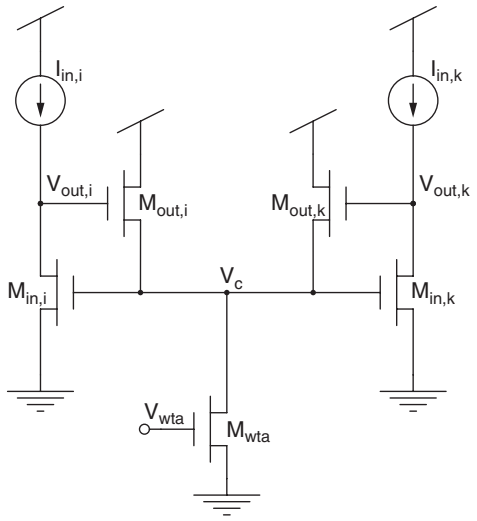


Figure 3.1: Classical current–mode WTA circuit: schematic diagram of a two node network.

duced to improve its performance, to arrive to the complete circuit implemented on the chip.

3.3 Current–mode WTA circuit description

The schematic diagram of a two node original "WTA" circuit as proposed in Lazzaro et al. (1989) is shown in Fig. 3.1. Two current conveyors receive two input currents $I_{in,i}$ and $I_{in,k}$ and compete for the bias current I_{wta} via the common node V_c. The bias current is generated by the NMOS transistor M_{wta}, operated in the weak inversion regime and in saturation. A complete analytical description of the two node WTA circuit is provided in Appendix B. Here I describe the circuit's behavior qualitatively.

If the two input currents are equal, the bias current is split equally between the two branches and the two output currents and voltages are equal. When one of the two input currents, e.g. $I_{in,i}$, increases with respect to the other, the two current conveyors begin to compete for the bias current, and the node receiving the highest input suppresses the other. Initially the current of the losing branch decreases linearly, due to the Early effect on the transistor $M_{in,k}$, then, for increasing difference between the two input currents, the input transistor of the losing pixel, $M_{in,k}$, is brought out of its saturation region and is shut off. In this case the output voltage of the losing pixel goes to Gnd, the output transistor $M_{out,k}$ is turned off, and the entire bias current flows through the winning branch.

The output of the WTA can be read either from the current signal as a binary variable which is high for the winner and low for all other nodes, or from the voltage output which also conveys the information about the magnitude of the winning input current. An analysis of the time response of the circuit is derived in Appendix B, together with a static analysis. There the stability condition for the network is derived as

$$I_{wta} > 4I_{in}\frac{C_{in}}{C_c}, \tag{3.1}$$

where I_{wta} is the bias current, I_{in} is the maximum input current, and C_{in} and C_c are the parasitic capacitors on the input and the common node respectively.

In the SAC I designed an explicit capacitance C_{in} at each pixel (C_{wta} in Fig. 3.2), in order to satisfy the stability condition. In this case the system exhibits first order dynamics:

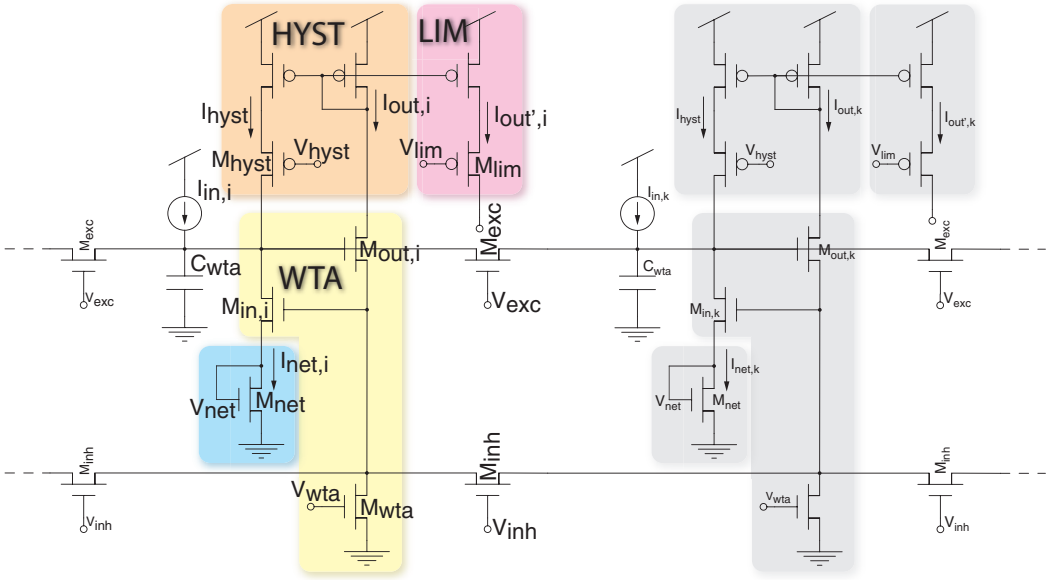


Figure 3.2: Current–mode hysteretic WTA circuit with diode–source degeneration, local excitation and inhibition.

the time constant of the winning cell is then $\tau_w = \frac{C_{in}k}{U_T I_{in}}$, and of the losing cells is $\tau_l = \frac{C_{in}V_e}{I_{in}}$, where $\frac{U_T}{k}$ ≈40mV and V_e is the Early voltage. Both time constants depend on the order of magnitude of the input currents. The resolution and speed of this network can be augmented by reducing the Early effect on the M_{in} transistors, as well as by diode–source degeneration of the input transistors (Lazzaro et al., 1989).

A schematic diagram of the enhanced network implemented on the SAC is shown in Fig. 3.2. The core circuit is shown in the "WTA" block. It comprises the current conveyor formed by transistors M_{in} and M_{out}, and the bias transistor M_{wta}. The block formed by the PMOS current mirror and transistor M_{hyst}, named "HYST", implements positive feedback and confers the hysteretic property to the competition. The second output of the current mirror ("LIM") is used as read–out of the competition. The other three transistors M_{net}, M_{inh}, and M_{exc}, implement diode–source degeneration, a diffusive network which sets global or local competition and a diffusive network which implements lateral excitation among neighboring cells, respectively. The gate voltage V_{net} of the diode connected transistor M_{net} is logarithmically proportional to the net input current of the corresponding WTA node, as given by $V_{net} = \frac{U_T}{k} \ln \frac{I_{net}}{I_0}$. I_{net} represents the sum of all currents converging on the input node, and is given by $I_{net} = I_{in} + I_{hyst} - I_{loc-exc}$, where I_{in} is the total input current, I_{hyst} is the positive feedback current from the hysteretic module, and $I_{loc-exc}$ is the amount of current going to the neighboring WTA nodes through the lateral transistors M_{exc}.

The positive feedback "HYST" block of Fig. 3.2 was introduced to further increase the resolution of the circuit and improve its speed (Starzyk and Fang, 1993). This way as soon as one cell begins to suppress the others and its output current increases, its input current also increases and the dynamics of the selection speeds up. The feedback current in the input node of the winning cell stabilizes the selection, and implements a form of hysteresis: a new cell can win the competition only if its input current exceeds the input current of the winner plus the positive feedback current I_{hyst}.

The transistor M_{hyst} of Fig. 3.2 was introduced to allow better control of the positive feedback current by decoupling it from the WTA bias current. A similar approach was proposed in Morris et al. (1996): instead of using the output current of the WTA coming

from the transistor $M_{out,i}$, they included a new NMOS transistor ($M_{out',i}$) with the gate connected to $V_{out,i}$ and the source connected to another NMOS (M_{hyst}), acting as current source; $M_{out',i}$ turns on only when the WTA cell wins the competition. Its output current, set by the current source, is then mirrored to the input node to implement hysteretic feedback, and sent to the output circuitry. This approach needs two more transistors to independently control the hysteretic feedback and the output current, but has the advantage of producing a current independent of the voltage $V_{out,i}$. With the method I have implemented, the hysteretic current is independent on the output voltage only when M_{hyst} is saturated. If the input current is too high, $V_{out,i}$ increases and the transistor M_{hyst} leaves saturation, turning off the positive feedback. This behavior has been verified experimentally in Sec. 4.3.2 and 4.3.5.

DeWeerth and Morris (1995) proposed to distribute the hysteretic current to neighboring cells via a diffusor network. This modification gives a competitive advantage to the pixels close to the winner and allows the network to select and track moving stimuli. In Fig. 3.2 the transistors M_{exc} implement a diffusor network applied to the input node of the WTA, diffusing both the hysteretic current and the total input current to the cell (Indiveri, 2001a). Besides being useful for tracking, the diffusion of the input current implements spatial smoothing, giving a competitive advantage to regions of activity, which in vision typically correspond to objects, compared with single pixels, and reducing the effect of mismatch between pixels. Indiveri (2001a) shows that the amount of current diffused to the $i-th$ cell, $I_{in,i}$, depends exponentially on the bias applied to the gate of the NMOS diffusors V_{exc}, and decreases with the increasing level of the input current $I_{in,k}$ of the stimulated WTA cell[3]:

$$I_{in,i} = I_{in,k} \left(\frac{I_0 e^{\frac{kV_{exc}}{U_T}}}{I_{in,k}} \right)^{|i-k|} . \tag{3.2}$$

The property of this particular diffusor network is that the spatial constant of lateral spread decreases with increasing input current levels, maintaining local lateral cooperation and avoiding smoothing over too a large region. In Sec. 4.3.4 I show experimental data demonstrating the effect on the competition of the lateral cooperation implemented through the diffusor network. Another diffusor network of the same type is applied to the common node of the "WTA" block (Indiveri, 2001a), as shown in Fig. 3.2. When it is enabled it limits the global WTA computation to a local region (Lazzaro et al., 1989): the winning cell inhibits only its neighbors, allowing multiple local winners in the array.

3.4 Conclusions

This chapter concerns the silicon implementation of interactions between basic elements that realize specific computations such as WTA competition.

I first introduced the computational relevance of a WTA network as an emergent property of recurrent networks, which processes information on the basis of context and is capable of extracting relevant information from noisy, ambiguous data. I then presented a compact and elegant circuit for the implementation of such a complex and crucial function. The current–mode WTA circuit is relevant in silicon implementations of cortical computation since it is very compact, is scalable, and operates at low power. Its most striking property is the simple connectivity among units, which allows the realization of dense bi–dimensional arrays with many units on a single chip. This approach with distributed inhibi-

[3] A full derivation of this equation is shown in Appendix B

tion is a compromise with respect to implementations with excitatory neurons and a single inhibitory unit (or population); such implementations (Hahnloser et al., 2000; Chicca et al., 2006a) have a richer behavior but are difficult to scale, especially in two dimensions, given the complex and area–consuming routing of signals from the excitatory to the inhibitory neurons and back.

I described the current–mode WTA circuit, reviewing the main contributions to its implementation by numerous authors, and presented the final version with the modifications I introduced for its implementation in the SAC. The WTA circuit I designed augments the computational capabilities of the core circuit first proposed by Lazzaro et al. (1989), by introducing diode–source degeneration, tunable hysteresis, lateral excitatory coupling, and local inhibition. My main contribution is in the introduction of the transistors M_{hyst} and M_{lim}, which allow better control over the hysteretic and output current respectively, over previously proposed solutions.

In the next chapters I characterize the SAC chip and show the role of the WTA circuit, as part of a more complex system with a specific function.

Chapter 4

The selective attention chip (SAC)

4.1 Introduction

The Selective Attention Chip (SAC) represents the "device" level of the neuromorphic quest, comprising the basic elements described in the previous chapters, to implement a specific functionality. It was designed to build in hardware selective attention architectures of the type proposed by Itti and Koch (2001). Specifically it receives a saliency map as input, and implements WTA competition and inhibition of return (IOR) to reproduce the scan–path of attention.

The SAC is interfaced to the external world via the AER communication protocol. It is designed to receive input spike trains that encode the saliency of the corresponding input stimulus in their mean frequency; it transmits the result of the computation by sending events to the external bus, the address of which corresponds to the pixel selected as the attentional target. The AER infrastructure confers great flexibility for the use of the SAC: its input can be generated via software simulation of computational models that extract the saliency map from images; alternatively, the SAC can process the activity of other AER neuromorphic chips. The SAC output can be further processed by AER chips such that it can be part of hierarchical multi–chip systems that model biological vision.

4.1.1 Relation to previous work

Previously proposed VLSI implementations of selective attention (Brajovic and Kanade, 1998; Horiuchi et al., 1997; Horiuchi and Koch, 1999; Indiveri, 1999; Indiveri et al., 2001, 2002) include photoreceptors and local but rudimentary saliency map computation circuits together with WTA and IOR, to implement single chip selective attention systems. The philosophy behind the SAC project was to separate the sensor and saliency map computation from the scan–path computation (Indiveri, 2000b), producing a more flexible tool for the exploration of different saliency map models. With this approach, based on the extensive use of AER communication, the SAC can process signals arising from multiple sensory modalities (e.g. visual and auditory). Additionally, the chip is designed with externally tunable parameters, that allow the IOR mechanism to be disabled and configure the local WTA competition. This flexibility and the use of the AER protocol allows the implementation of hierarchical systems, where for example multiple instances of the SAC could be used to implement center–surround inhibition and the normalization required for feature map generation in the Itti and Koch (2001) model.

The SAC is the evolution of preceding designs proposed in (Indiveri, 2000a,b). It consists of a 32×32 array of computation units and additional features and modifications to the

basic circuits.

The new features I introduced are short–term depression (STD) in the input synapses (Sec. 2.4.1) and spike frequency adaptation in the output neurons. STD adapts the weight of a synapse with an increasing number of input spikes. Its function is to decrease sensitivity to constant stimuli, while enhancing their changes in time (Sec. 4.3.5). Spike frequency adaptation decreases the spiking activity of the I&F neuron for an increasing number of spikes emitted, therefore it reduces the number of events sent to the external bus (Sec. 4.2.4).

The input synapses of the SAC which convert their input spike trains into currents for the input of the corresponding WTA circuit, and the inhibitory synapse used to implement the IOR mechanism, are implemented with the new DPI circuit (Sec. 2.2.6). The DPI generates smoother currents compared with previously proposed circuits, such as the CMI (see Sec. 2.2.4 and Sec. 2.2.6), which consequently renders the WTA circuit less sensitive to the timing of individual input spikes.

I added circuits to give better control over the feedback hysteretic current and the output current of the WTA pixel (Sec. 4.2.3). Finally I developed circuits for interfacing the internal computational core of the chip to the output: I introduced new AER circuits for decoding, arbitration, and handshake, and I designed a decoder circuit ("Select" of Fig. 4.1) together with source follower circuits for reading an internal voltage that allows monitoring of the state of any pixel of the array (Sec. 4.2.1).

The SAC presented in this work is the third prototype generation: A first prototype was fabricated still using the CMI as input synapse, but with new interfacing circuits for the AER communication protocol. This implementation unveiled the inadequacy of the CMI synapse: the winner of the WTA competition depended mainly on the timing of each individual input spike, rather than on their mean frequency. A second prototype was implemented with the reset and discharge synapse described in Sec. 2.2.2; this synapse integrates input spikes linearly until it saturates to the maximum output current, independent of the input frequency. For this reason the use of this prototype was impractical. The third and final implementation uses the DPI for the input synapses. The results, described in the next sections, demonstrate that this SAC prototype can be reliably used for modeling the mechanism of attentional scan–path generation.

In the following sections of this chapter I describe the SAC architecture and schematics, then characterize its functionality. In Chap. 5 I show the behavior of a multi–chip system which uses the SAC in the context of selective attention modeling.

4.2 The chip's architecture

The SAC was fabricated in standard AMS 0.35μm CMOS technology, via the Europractice IC service (*http://www.europractice.imec.be/europractice/europractice.html*).

Fig. 4.1 shows the layout of the SAC. The core of the chip comprises an array of 32×32 pixels, each one is $90\times45\mu m^2$, and the whole chip with external interfacing circuits and pads occupies an area of $10mm^2$.

The SAC has been designed to be one of the processing stages of a multi–chip AER system, receiving as input the activity of a neuromorphic sensor, e.g. a silicon retina and/or silicon cochlea, and producing an output which can be sent to an actuator, or to higher processing stages implemented by other neuromorphic devices. This has been possible thanks to the communication system adopted, the Address–Event Representation (AER), which since its first formulation became a standard communication protocol used by neuromorphic chips. As mentioned in the introduction, a successful approach in the design of neu-

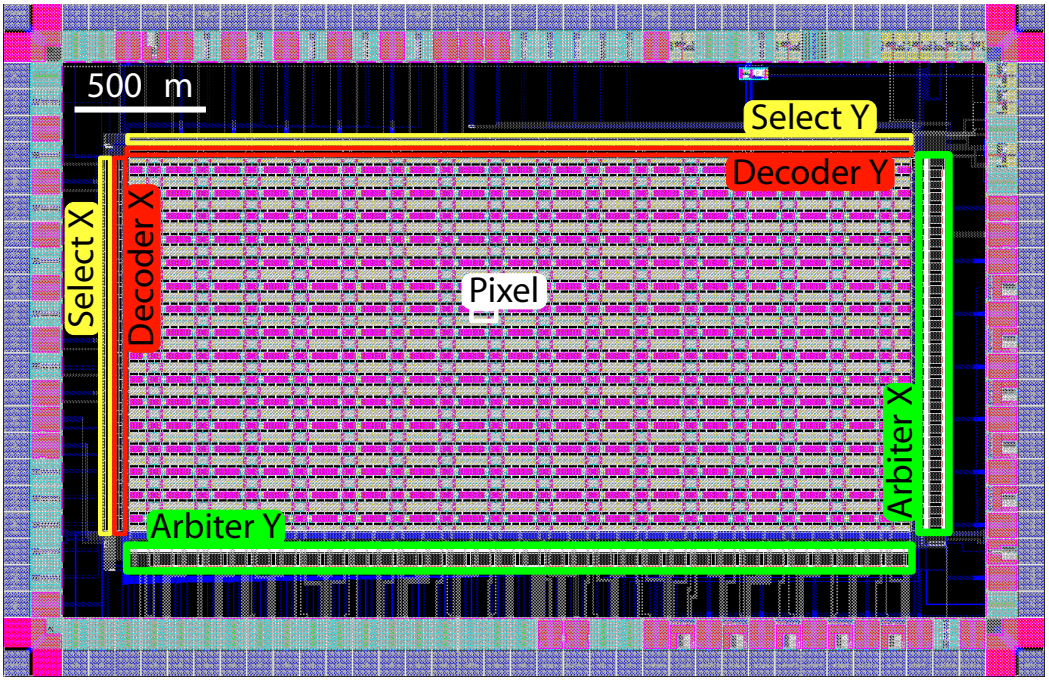


Figure 4.1: Layout of the SAC: the outer ring comprises all the pads connecting the internal wires to the pins of the package, accessible from outside; The internal "Core" comprises the 32×32 pixels performing the computation; the surrounding circuitry, "Decoder X, Y", "Arbiter X, Y", performs the connection to the external AER bus. "Select X, Y" is an additional decoder used to route the V_{net} node (see Fig. 3.2) of the addressed pixel to an output pad, for testing purposes.

romorphic system is to rely on both analog computation and digital communication. The analog computation is performed in the core array of the SAC through the interaction of excitatory and inhibitory synapses, neurons, and the WTA network. The digital communication is accomplished by the transmission of address events, based on the AER. The "Decoder" and "Arbiter" blocks in Fig. 4.1 interface the analog core with the external AER bus. The "Select" block comprises an additional decoder for selecting a chosen pixel and monitoring an internal variable (V_{net} of Fig. 3.2) that conveys information about the state of the pixel.

Additionally the analog circuits have tunable parameters that, set via external voltage references, can change the properties of the core computation.

Fig. 4.2 shows the block diagram of a SAC pixel; each pixel in the array comprises an input excitatory synapse which translates its input spikes into the current I_{exc}. A current–mode hysteretic WTA competitive cell compares the input currents of each pixel; the winning cell sources a constant current to the corresponding output leaky Integrate and Fire (I&F) neuron (Indiveri, 2003b). The identity of the spiking neuron signals which pixel is winning the competition for saliency, and therefore the pixel that had received the highest spiking input frequency. The output spikes of the I&F neuron are also sent to a feedback inhibitory synapse, which subtracts current (I_{inh}) from the input node of the WTA cell; the net input current to the winner pixel decreases, and a new pixel is eventually selected. This self–inhibition mechanism is known as Inhibition of Return (IOR) and allows the network to sequentially select the most salient regions of input images, producing the attentional scan path (Itti and Koch, 2001).

In the following paragraph I describe the blocks listed above, their functionality and their significance within the context of the computation performed by the SAC. In the sec-

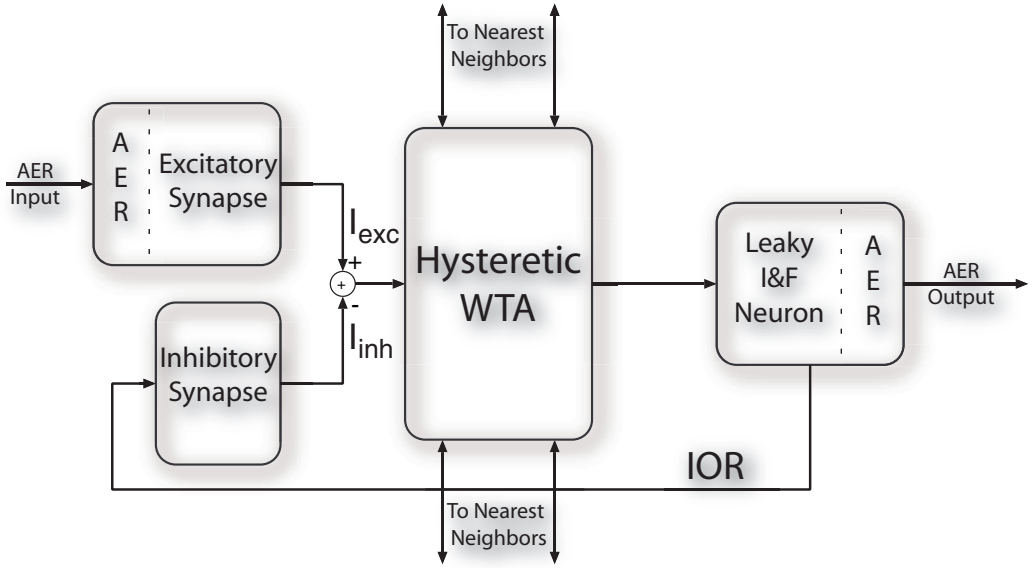


Figure 4.2: Block diagram of the SAC pixel: input AER spikes are converted into the current I_{exc}, the WTA cells compare these currents, only the cell winning the competition sources an output current into the membrane of the corresponding read–out I&F neuron. The address of the active neuron, sent to the AER bus, signals the pixel selected for attentional deployment. The spikes of the neuron are also integrated by the local inhibitory synapse, that subtracts the I_{inh} current from the WTA input, thus implementing IOR.

ond part of this chapter I describe the experiments performed to characterize the chip, especially as a function of its most significant parameters that can be tuned to modify the network behavior.

4.2.1 The Address–Event Representation

Fig. 4.3 shows an outline of the AER communication strategy (Lazzaro et al., 1993). The AER communication protocol is event–driven: each event is a digital pulse, with stereotyped height and width. For each event the address of the corresponding neuron is sent to the bus, and information is encoded in the temporal pattern of the events. Time is self–encoded, and does not need to be explicitly transmitted; when an AER bus is used to connect two neuromorphic chips with the same communication latency, the sequence and the implicit relative timing structure of the events are sufficient. Only when the communication happens between a neuromorphic chip and a different device (e.g. a computer) with an intrinsically different information encoding scheme, the device can append a time stamp to the address of each event. The temporal accuracy of the system is determined by the minimum communication cycle duration, which is typically much faster than the refractory period of a single neuron.

A detailed analysis of the possible approaches in communication for neuromorphic chip and their implementation issues is presented in Boahen (2000). Specifically, in the SAC an arbitrated and pipelined version of the protocol is used: coincident events are arbitrated to gain access to the external bus.

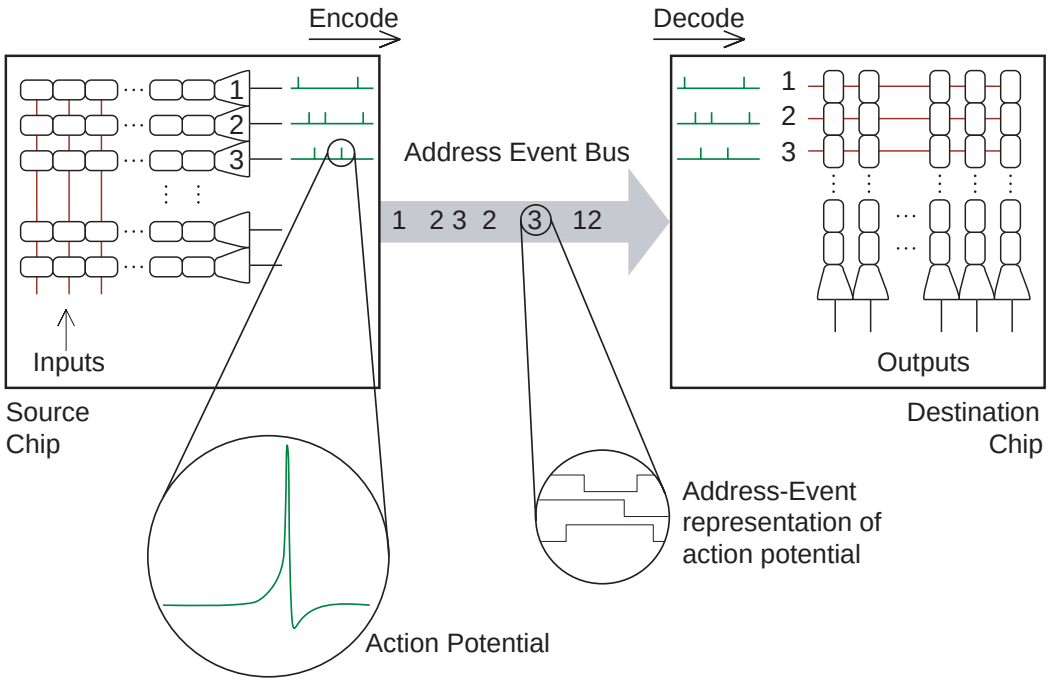


Figure 4.3: Schematic diagram of the AER communication scheme: the first neuron that emits a spike writes its address on the bus, the address is decoded by the receiver and sent to the target synapse.

AER protocols

There are several different standards for AER protocols; here I describe the two protocols used in the SAC: the point–to–point (P2P) and the silicon–cortex (SCX).

The P2P protocol is used for the communication from one chip, the sender, to another chip, the receiver, via a dedicated bus. AER0.02 (AER) is the AER P2P standard protocol first proposed and formalized, extending the concepts in Lazzaro et al. (1993). The SCX protocol (Douglas et al., 1994; Deiss et al., 1998) is a multi–sender, multi–receiver extension of the AER0.02, which allows communication between many chips sharing the same physical bus. Fig. 4.4 shows the time diagrams of both protocols.

The SAC is a transciever, i.e. it receives and sends spikes; the protocol adopted for the input side is P2P AER0.02. For the output side it uses the SCX protocol. It uses a parallel bus for transmitting the 10 address bits (5 for the 32 rows and 5 for the 32 columns) along with request and acknowledge lines, required to handshake with the communication devices. In the next paragraphs I describe in detail the time flow of the signals involved in the two protocols, as shown in Fig. 4.4; for the P2P I include the internal signaling of the receiver side, and for the SCX the sender side, as they are implemented on the SAC.

- **P2P protocol:** In the P2P protocol two chips communicate via a dedicated bus. When the sender is ready to send an event, it first writes the address on the bus data lines then sends a request "Sender Req" signal to the receiver. The receiver confirms that it received the event with an acknowledge "Receiver Ack", then the sender can remove the request and start a new arbitration sequence. Inside the receiver chip, the address on the bus is decoded, and the arrival of the "Sender Req" signal triggers an internal request, "Synapse Req". At the transition of the "Synapse Req" the transmission of a digital pulse to the synapse corresponding to the decoded address is enabled. When the pulse reaches the synapse circuit, an internal acknowledge, "∼Synapse

Ack", is triggered. "Synapse Req" and "Receiver Ack" are reset to inactive state, as is "∼Synapse Ack". Only when the receiver removes its acknowledge can the sender send a new address with a new request, and only when the internal synapse acknowledge is inactive can the new event be routed to the corresponding synapse. If the "∼Synapse Ack" reset is too slow compared with the generation of a new request from the sender, communication is delayed. This effect is observed in a longer duration of the successive "Sender Req" (red transition). The red circle indicates an erroneous address change during the synaptic stimulation; it does not affect the computation, thanks to latch circuits that fix the data lines during synaptic stimulation.

- **SCX protocol:** The SCX protocol is used when a bus is shared; the sender writes on the bus only when it receives an acknowledge signal indicating that the bus is free, otherwise the data lines are set in high impedance, and other senders can drive them. When the sender is ready to send an event, it generates a request signal "∼Sender Req". If the receiver activates the "∼Receiver Ack", the sender writes the address of the spiking neuron to the bus data lines. When the internal routing of the spike is completed, the receiver resets the "∼Receiver Ack" and the sender releases the bus, setting the drivers in "High–Impedance" state, such that any other device sharing the bus can start its communication cycle and drive the address lines. Arbitration occurs inside the sender chip, to determine which event can gain access to the bus. When a neuron spikes, it sets "∼Neur Req X" low, and arbitration between the rows takes place. When the arbitration completes, a "∼Neur Ack X" signal sets "∼Neur Req Y" low. The arbitration between columns generates the output request sent to the receiver "∼Sender Req", the same signal latches the address bits of the selected neuron on the output pads. The "∼Receiver Ack" enables the pads to drive the bus data lines, and is sent to the arbitration circuits, which in turn generate the "∼Neur Ack Y" signal, resetting the neuron. The"Neur Req" signals are then set back to inactive state. The reset of "∼Receiver Ack" disables the output data lines and resets the internal "Neur Ack" signals, then a new cycle can start.

In the next paragraphs I describe the digital circuits connecting the analog core of the SAC with the external AER bus, highlighting the modifications introduced with respect to the circuits previously used in our institute.

SAC AER input

The input of the SAC is designed to receive P2P based data. Its input circuitry decodes the address events, sends the spike to the corresponding synapse, and acknowledges the sender. The decoding is performed by the "Decoders" of Fig. 4.1, the handshaking is implemented by a "C–element" (Shams et al., 1998). This asynchronous circuit waits for the internal acknowledge "Synapse Ack" from the synapse (see Fig. 4.4(a)) before acknowledging the sender. When the internal dynamics of the spike routing is slow with respect to the event generation from the sender, communication is delayed and the sender waits before sending a new event. The internal acknowledge node is connected to all of the 32×32 synapses by a wired–OR circuit. To reduce the capacitive load of this node, and speed–up the internal dynamics of the spike routing, I implemented a wired–OR between the acknowledge line of each row. Fig. 4.4(a) shows the time diagram of the P2P protocol, together with the internal handshake of the chip. The specific implementation on the SAC is based on a temporal assumption: namely that the data lines are ready and stable before the request arrives, for

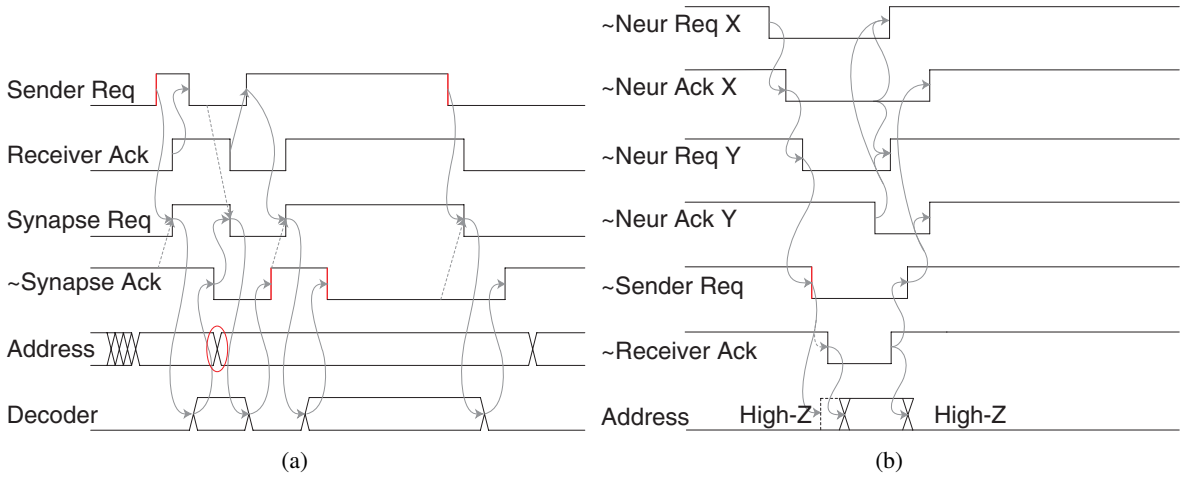


Figure 4.4: Time diagrams of P2P and SCX protocols, together with the internal signals that coordinate the activity of each neuron and synapse with the external communication. P2P is used at the input stage of the SAC, SCX at the output. (a) P2P protocol (active high): the bus is dedicated, the sender first writes the data on the bus, then sends a request (data ready) to the receiver. The communication cycle is concluded when the receiver acknowledges (data red) the sender. The address lines are latched during the internal routing to the input synapses, such that a fluctuation of the data does not affect the transmission of the event. The red circle shows an erroneous address change during the synaptic stimulation. (b) SCX protocol (active low): the bus is shared, each sender writes on the data lines only when the bus is free. Otherwise the bus is in high impedance state. The symbol "∼" before the name of a signal indicates that its active state corresponds to a logical 0.

the data to be decoded before they are latched by the internal request. This assumption is fulfilled by the sender timing specification.

SAC AER output

The AER output circuits of the chip arbitrate the events coming from the neuronal array and handle the communication with the external AER bus. Fig. 4.4(b) depicts the time diagram of the signals involved in the communication protocol, showing both signals exchanged with the external bus, and internal ones from and to the neurons. Spikes generated by the neurons are reset by the internal "Neur Ack" signals from both the X and Y arbiter, confirming that the spike was arbitrated, sent to the bus, and processed by the receiver. The arbiters handle spike collisions, and encode the address bits that are sent to the external bus. The arbiter implemented on the SAC is a non–greedy arbiter (Boahen, 2004).

Once the events are arbitrated, the selected neuron's address is encoded and sent to the output pads. The data, however, can only be written on the external bus when the "Receiver Ack" signal is active.

I designed a newer version of the output pads (PAD_LATCH shown in Fig. 4.5) for the accomplishment of this specification: the "∼Sender Req" generated by the arbiter is used as triggering signal for latching the encoded address bits, then the "∼Receiver Ack" signal enables a tri–state buffer to drive the bus lines; when the acknowledge signal is removed by the receiver, the tri–state sets its output in High–Impedance state, and the bus can be driven by an other device.

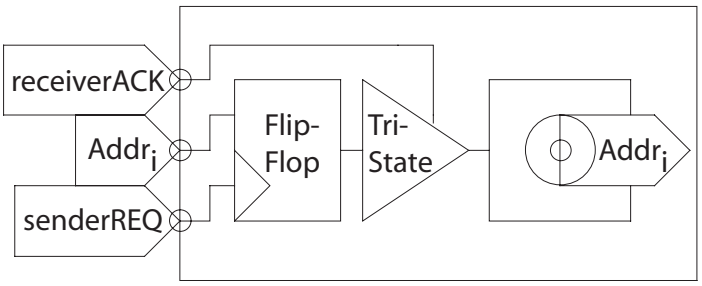


Figure 4.5: PAD_LATCH: diagram of the output pads implemented on the SAC chip. The address bit (Addr$_i$) is latched by a flip–flop when the request signal is generated from the arbiter (senderREQ); the data is ready at the input of a tri–state buffer, enabled by the acknowledge signal from the receiver (receiverACK). The tri–state leaves the bus lines in High–Impedance state, free to be driven by another device.

4.2.2 The input excitatory synapse

The input spikes decoded from the AER bus are routed to the excitatory input synapse of the corresponding pixel in the array. The synapse circuit implemented in the SAC is a simplified version of the synaptic circuit presented in Chap. 2. It comprises the core DPI circuit and the short–term depression circuit, with the exclusion of the "NMDA" and "G" functionality, as shown in Fig. 4.6. This synapse translates digital input pulses into output currents with an exponential time course, modeling the behavior of biological synapses, as described in Chap. 2. In the specific application of the SAC, the synapse has to generate a current proportional to its input activity, representing the salience of the correspondent region of the stimulus. I exploit the property of the synapse expressed by eq. (2.26): the mean output current of the DPI in response to a train of spikes with constant frequency is linearly proportional to the input frequency, therefore the input current to each WTA cell is proportional to the saliency of the correspondent region of the stimulus. Besides the use of the DPI circuit, the implementation of short–term depression is the major innovation introduced in this version of the SAC with respect to previous prototypes (see Sec. 2.4.1 and 4.3.5).

4.2.3 The WTA circuit

The computational core of the SAC is the hysteretic WTA circuit described in Sec. 3.3. Each WTA pixel receives its input current from the correspondent excitatory input synapse. It selects the pixel receiving the highest input current, while suppressing all the others.

The particular circuit implemented is shown in Fig. 3.2. It allows a dual read–out: the output voltage logarithmically encodes the input value of the maximum current; the output current is a binary value: when high it signals that the cell is winning the competition, when low it signals that the cell is not winning.

For the scope of a selective attention implementation, the information conveyed by the current, which identifies the position of the winner, is sufficient. This current can be read–out by a second output transistor in the current mirror, implementing positive feedback, and a limiting transistor in the "LIM" block of Fig. 3.2, with the function of decoupling the output current from the biasing current of the WTA.

To monitor the state of each WTA pixel it is also possible to read the output voltage V_{net} from the gate of the diode–connected transistor M_{net} of Fig. 3.2. In the SAC, each V_{net} is connected to a source follower; a decoder ("Select" in Fig. 4.1) enables a single WTA pixel to access an output pad, which can then be used to monitor the activity of the chosen

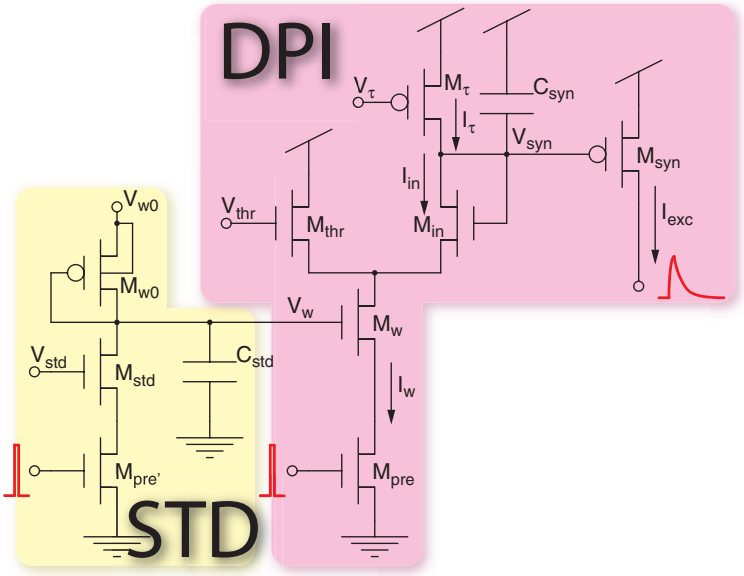


Figure 4.6: Circuit diagram of the excitatory synapse implemented on the SAC. It comprises the DPI circuit and the STD circuit described in Chap. 2

WTA cell.

4.2.4 The output Integrate & Fire neuron

The output current of each WTA cell is transformed into a train of pulses by a read–out Integrate and Fire (I&F) neuron. We use a neuron to convert the output current into a train of pulses in order to exploit the AER communication system at the output of the chip. It is a convenient and robust method for multiplexing the activity of large networks on a single bus, and is necessary for including the SAC in multi–chip systems, where for example we can stack multiple instances of the SAC to build hierarchical models of selective attention (see Sec. 4.2.1). In this scheme the address of the active neuron encodes the result of competition and, in the case of visual selective attention, the position in the visual field at which to deploy attention.

The I&F neuron circuit adopted for the SAC is a phenomenological model of a spiking neuron; the rationale for this choice is described in the introduction. The circuit implemented on the SAC, namely the low–power leaky adaptive I&F neuron model, has been proposed and fully characterized by Indiveri (2003b). It is the result of the evolution of previous I&F circuits, starting from the "Axon–Hillock" circuit, proposed by Mead (1989); the modifications introduced since that first implementation are crucial for improving its compactness and power consumption, and include additional features typical of biological neurons, such as spike frequency adaptation.

Circuit description

The I&F circuit implemented on the SAC (Indiveri, 2003b) merges the design principles of previously proposed circuits (Mead, 1989; Schultz and Jabri, 1995; Boahen, 1997; van Schaik, 2001; Rasche, 2005), satisfying the conflicting needs for compactness and low power with flexibility and richness of behavior. Fig. 4.7 shows the diagram of the circuit; it can be divided into six blocks, plus the membrane capacitor. The "AER" block implements

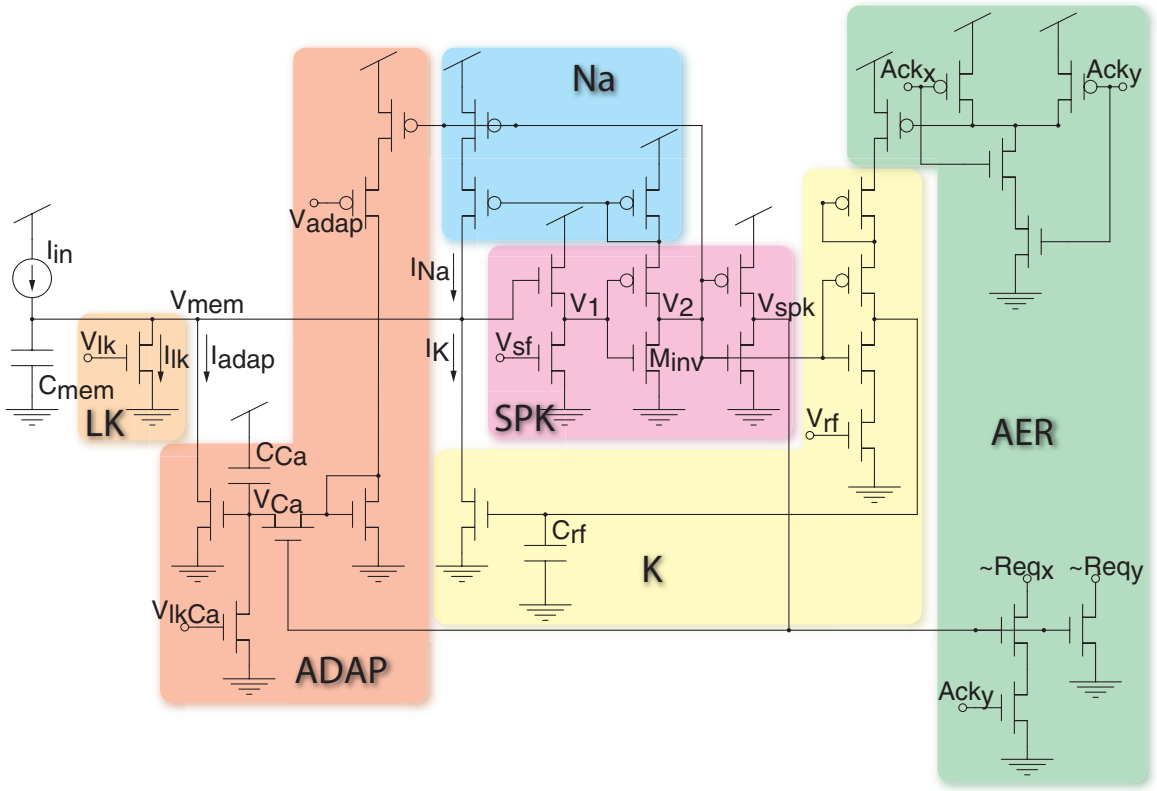


Figure 4.7: Integrate–and–fire circuit diagram: The input current (from the WTA cell) is integrated by the membrane capacitor C_{mem}. When the membrane potential rises above a threshold set by the tunable bias V_{sf}, the source follower in the "SPK" block gets activated and drives the first inverter, connected to the positive feedback ("Na"); it injects a positive current onto the membrane further increasing V_{mem} and speeding up the spike generation. The second inverter in the "SPK" block generates the fast digital spike sent to the arbitration circuits that manages the neurons' access to the external bus. The "K" block mimics the resetting and hyperpolarizing function of the late potassium channels; it is activated when communication with the AER bus is completed and subtracts current from the input node, resetting the membrane potential to its resting value; the tunable dynamics set by the starved inverter, via the voltage V_{rf} and the capacitor C_{rf} set the refractory period of the neuron, i.e. the time interval after the spike emission during which the neuron is not responsive to any input current. The transistors of the block "ADAP" integrate the spiking activity of the neuron on a voltage variable representing the internal calcium concentration of the neuron; after a period of prolonged activity a current is subtracted from the neuron's input node, decreasing its spiking activity. The transistor "LK" implements a constant leak from the membrane. The block "AER" handles the connectivity with the x and y arbitration circuits; in particular the neuron cannot be reset unless the acknowledge (Ack_x, Ack_y) signals return in response to the requests (Req_x, Req_y) sent to the arbitration.

the connection with the arbitration circuits (see Sec. 4.2.1). The "LK" block implements the constant current leak from the neuron's membrane, "SPK" detects when the membrane voltage crosses the (tunable) spiking threshold and produces the output digital spike, "Na" implements the positive feedback of the spike generation, injecting a positive current into the membrane, and macroscopically reproducing the irreversible activation of voltage sensitive sodium channels. The sodium current is active only during the spike and is inactive otherwise. The current generated by the "K" block resets the membrane voltage to its resting level when the communication cycle with the external bus has finished, and sets the refractory period duration. The "ADAP" block implements the spike frequency adaptation mechanism.

When a constant input current, I_{in}, is injected into the membrane capacitor the mem-

brane potential linearly increases, and the source follower in the block "SPK" driven by V_{mem} generates the signal $V_1 = \kappa(V_{mem} - V_{sf})$; V_{sf} sets the neuron's threshold voltage. As V_{mem} increases and V_1 approaches the threshold voltage of the transistor M_{inv}, V_2 starts to decrease and the current through the current mirror of the "Na" block starts to increase. As a result the current I_{Na} further increases V_{mem} and V_1, implementing a positive feedback mechanism that has the effect of making the inverters switch very rapidly, dramatically reducing power consumption. When the output voltage of the first inverter in the block "SPK", V_2, is sufficiently low to drive the second inverter, the digital spike is generated and the communication cycle with the AER arbiter starts. When the spike is sent to the output bus and the acknowledge signals arrive both from the X and Y arbiters, the starved inverter in the "K" block turns on and starts discharging the membrane capacitor C_{mem}. The tunable bias V_{rf} sets the slew rate of the inverter and therefore the rise of V_k, controlling the re–polarizing current I_k. This current drives V_{mem} to ground and keeps it clamped there for a period dependent on V_{rf}. Different values of V_{rf} set different durations of the refractory period, which in turn sets the maximum possible firing rate of the neuron.

During spike emission the "ADAP" block is also active; the PMOS current mirror sources a maximum current, as set by the tunable voltage V_{adap}, charging the capacitor C_{Ca}; the voltage V_{Ca} increases with every spike, and during the Inter–Spike Interval (ISI) it leaks to ground through the transistor M_{lkCa}. This voltage represents the intracellular calcium (Ca^{2+}) concentration, related to the spiking activity of the neuron. As V_{Ca} increases, a negative current I_{adap} exponentially proportional to V_{Ca} is subtracted from the input, reducing the spiking frequency of the neuron.

The adaptation mechanism used in the SAC reduces the activity of the neurons. In the SAC application, a constant output firing rate at steady state is not informative. The adaptation mechanism helps to decrease the number of events produced by each single neuron, reducing the traffic on the AER bus, hence reducing bandwidth use and power consumption.

Besides being used for monitoring the activity of the neurons to follow the movement of the focus of attention, the output spikes of each neuron are integrated by the corresponding local inhibitory synapse, which subtracts the current I_{inh} from the input of the WTA node, implementing a self–inhibition mechanism that allows the WTA network to deselect the current winner and select a new one.

4.2.5 The inhibitory synapse

The inhibitory synapse provides a mechanism for deselection of the winning neuron, in favor of stimulus exploration. Hardware systems based on the concept of selective attention and WTA competition that lack such deselection mechanism have been specifically designed for tracking a selected target (Brajovic and Kanade, 1998; Fish et al., 2004). In such implementations, once a target is selected, the system is designed to lock on the target, disregarding distracters and even new salient stimuli.

Horiuchi et al. (1997) proposed a system implementing attentional tracking, based on a different strategy: the authors use the direction of motion of the selected target for implementing a form of smooth pursuit; furthermore, they give a competitive advantage to the selected target and its neighbors by using the hysteretic mechanism described in Sec. 3.3. This system is very robust for tracking: for high values of the hysteretic current the system does not select any stimulus other than the selected target. For lower values of the positive feedback current, other stimuli with strong salience are able to win the competition and cause shifts of attention.

In the SAC I implemented both hysteresis, which favors tracking, and a self inhibitory mechanism, which favors shifts of attention. The dynamic interplay of these two mechanisms creates a complex behavior, mimicking the rich mixture of attentional tracking and shifts of natural scan paths. Similar systems have been proposed in the context of aVLSI implementation of visual selective attention (Indiveri, 2001b; Horiuchi and Niebur, 1999). In the next paragraph I briefly review the inhibitory mechanism implemented in such systems, and describe the one implemented on the SAC.

Circuit description

VLSI devices that include WTA competition and distributed hysteresis, together with an inhibitory mechanism, have been proposed in Horiuchi and Niebur (1999); Morris et al. (1996). In both cases the winning cell activates a local inhibitory circuit that subtracts current from the input node of the corresponding WTA cell. This allows the network to select the second strongest input; when the second winner is also self–inhibited the WTA starts the competition again; depending on the time decay of the inhibitory current the first winner can be selected again, or can continue to be inhibited, allowing the network to choose the third most salient stimulus. The inputs are scanned in order of saliency as long as the inhibition of the first winner is active.

Both architectures proposed by Morris et al. (1996); Horiuchi and Niebur (1999) implement self–inhibition by integrating the analog output voltage of the WTA over time using a low–pass filter circuit. In the SAC, the WTA output is transformed into spiking activity of the I&F neuron; this allows us to use the DPI synapse circuit as a temporal integrator and obtain a more flexible and tunable inhibition mechanism. The inhibitory version of the DPI synapse integrates the spikes of the winning I&F neuron and generates an inhibitory current I_{inh} which is subtracted from the input current I_{exc} (see Fig. 4.2). Fig. 4.8 shows the circuit diagram of the inhibitory synapse. The time course of build up and decay of inhibition can be tuned by changing the synaptic weights V_{winh} and V_{thrinh}, the synaptic time constant $V_{\tau inh}$, and by modulating the firing activity of the I&F neuron.

A long decay of inhibition prevents the system from choosing previously selected targets; such a behavior has been observed in psychophysics experiments and is referred to as Inhibition–of–Return (IOR) (Posner, 1980).

The scan path of the network depends on the IOR settings, on the relative magnitudes of the input currents, and on the hysteretic current. The number of selected neurons in the scan path increases with the duration of the inhibition decay; the higher the hysteretic current, the longer the duration of the active period of each neuron.

The SAC implementation of the IOR mechanism is similar to the software implementation in the Itti and Koch (2000) model. In both cases a negative input is added to the saliency map as soon as one pixel is selected as the attentional target. In the software model this input is a difference of gaussians (DOG), with a negative peak centered at the winning pixel and positive flanks at a given distance. This implementation of IOR accounts for the observed bias for making short saccades close to the current fixation point (Parkhurst et al., 2002). In the SAC implementation, the resistive network at the input of the hysteretic WTA described in Sec. 3.3 implements a lateral diffusion of the inhibitory current, with a negative peak at the center which decays exponentially with distance. Given the passive nature of this resistive network, there are no positive flanks in this model.

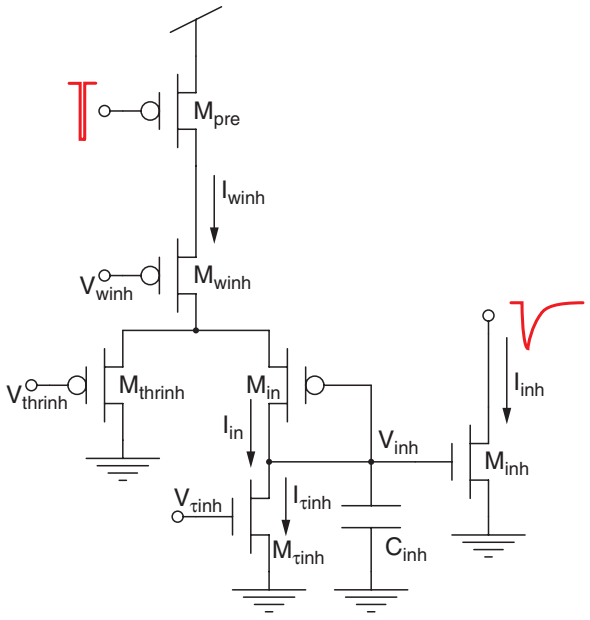


Figure 4.8: Circuit diagram of the inhibitory synapse implementing the IOR mechanism. The circuit is the complementary version of the DPI circuit (see Fig. 4.6); it integrates the spikes coming from the I&F circuit, generating the I_{inh} current which is subtracted from the input node of the corresponding WTA cell.

4.3 SAC functional characterization

In this section, I describe the results of experiments performed on the SAC to characterize the different circuits and their role in scan–path computation. I payed particular attention to the effect of the bias voltages that influence the overall behavior by modifying the competition, the lateral cooperation and the dynamics of the scan–path. Tab. 4.1 lists the biases of the pixel's circuits, and their main role. Many of them interact in a non–linear way and produce a rich variety of behaviors.

For each experiment all biases were set to reasonable values, then one or more biases were swept while monitoring the behavior of the system. With each experiment I explored the parametric space and determined the optional values of the biases for the application, then I used these values for subsequent experiments.

Methods

The experimental setup used to characterize the SAC comprises many different hardware and software elements: the SAC, a PCI–AER board, an oscilloscope, digital to analog converter (DAC) boards and a Linux desktop, as shown in Fig. 4.9 (see Appendix C for a description of the full multi–chip system setup). The DAC boards were used to automatically explore the parametric space of the chip bias voltages, and were controlled via Matlab commands from the Linux workstation.

The PCI–AER board interfaced the SAC to the Linux workstation; it was used to send spike trains to the SAC and to monitor its spiking activity.

Via the oscilloscope I monitored internal voltages that convey information about the status of the chip; in particular pixel number (0,31) of the array is used as test pixel: its voltages V_{syn}, V_w, V_{mem} and V_{inh} (see Figs. 4.6, 4.7, and 4.8 respectively) are connected to pads and can be monitored to characterize the behavior of the different circuits. The

Excitatory Synapse Circuit	
V_{w0}	synaptic weight (initial value)
V_{wstd}	short term depression strength, sets the dynamics and the minimum value of the depressed weight
V_{thr}	synaptic weight (non depressing)
V_{τ}	synaptic time constant
WTA Circuit	
V_{wta}	WTA bias current
V_{inh}	lateral inhibitory connections, controls the spatial constant of diffusion
V_{exc}	lateral excitatory connections, controls the spatial constant of diffusion
V_{hyst}	maximum value of the hysteretic feedback current
V_{lim}	maximum value of the current injected into the I&F circuit
I&F Circuit	
V_{lk}	constant current leak
V_{sf}	spiking threshold
V_{rf}	refractory period, saturation frequency
V_{adap}	spike frequency adaptation weight
V_{lkCa}	spike frequency adaptation calcium leak
Inhibitory Synapse Circuit	
V_{winh}	synaptic weight
$V_{\tau inh}$	synaptic time constant
V_{thrinh}	synaptic weight

Table 4.1: Description of the SAC circuit biases (excluding the biases of the external circuitry for AER and testing structures) with short description of their main role. In the next sections I explore the parametric space created by these variables, to characterize their functional role.

"Select" circuitry described in Sec. 4.2 can be used to monitor the voltage V_{net} of any pixel in the array (see Fig. 3.2).

The measurements made to characterize individual circuits, such as the excitatory synapse, the STD and the IOR synapse were performed on the test pixel (0,31). The experiment characterizing the effect of biases on the competition are always performed by stimulating pixels (10,10) and (15,10), named pixel 1 and pixel 2 respectively in the following sections. With this approach one can observe the relative changes introduced by different parameters on the same system, without introducing effects due to mismatch. The two pixels were chosen randomly to represent the typical activity of the network.

4.3.1 Mismatch evaluation

A first step in the characterization of the chip involves the evaluation of the effect of intrinsic inhomogeneities of the silicon implementation, known as mismatch. Mismatch is intrinsic in VLSI chips and arises mainly from defects in silicon doping; there are design techniques that help reduce the mismatch effect (Liu et al., 2002), such as increasing the size of critical transistors and designing all transistors in the same orientation these techniques were applied during the design of the SAC; in particular the MOSFETs M_w, M_τ of Fig. 4.6,

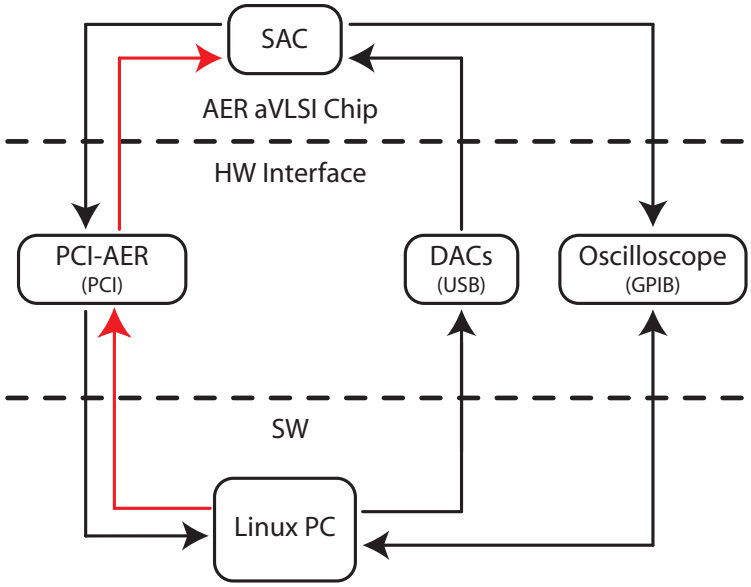


Figure 4.9: Experimental setup for the functional characterization of the SAC: the SAC activity is monitored via the oscilloscope and the PCI–AER board that is also used to generate the input to the SAC. DAC boards are used to set the chip's parameters and a Linux desktop coordinates the different parts of the system.

M_{winh}, $M_{\tau inh}$ of Fig. 4.8, and M_{wta} of Fig. 3.2 have a large gate area. I tested the effect of mismatch on the input of the WTA circuit, and on the output frequency of each neuron, by stimulating each pixel separately with a constant input mean[1] frequency of 100Hz, and recording the data under four different conditions: without competition, with competition, with competition and hysteresis, and finally adding a limitation on the current injected into the I&F neuron. In each condition the effect of mismatch was measured for different values of V_{wta}, the bias of the WTA network. For each pixel I acquired the oscilloscope trace of V_{net} for 200ms and the output frequency for 2s and computed its mean over time, then I calculated the mean (m) and standard deviation (σ) of this value over the pixels array. The measure reported in the following text for evaluating the relative effect of mismatch on the array is the coefficient of variation, $CoV = \frac{\sigma}{m} \cdot 100$.

As the WTA circuit decouples the input from the output currents, the mismatch between currents at the input node does not contribute to the mismatch on the output currents. The variation among the input currents of the WTA pixels depends on mismatch in the excitatory synapse circuits, and on the differences in the current mirrors that source the hysteretic current. The coefficient of variation of the input currents of the WTA pixels varies in a range between 0.61% and 0.7%, depending on the circuit settings. The mismatch effect on the input synapses results in a competitive advantage of some pixels over others.

The main potential source of mismatch that has an effect on the output of the SAC arises from the biasing transistors M_{wta} in each pixel (see Fig. 3.2). Fig. 4.10 shows the CoV on the output frequency of each pixel for the case in which the WTA competition is disabled ($V_{inh} = Gnd$ in Fig. 3.2). In such a case, all M_{wta} transistors are decoupled and each output neuron is affected by mismatch. If the WTA competition is fully enabled ($V_{inh} = V_{dd}$) all M_{wta} transistors operate in parallel, and the current sourced into the I&F neuron by the winning WTA cell is the same everywhere.

Mismatch can also arise from the current mirror and from the limiting transistor M_{lim}

[1]The standard deviation of the frequency, measured by monitoring the PCI–AER board sequencer output, is extremely small ($\sigma_{f_{in}} = 4e^{-13}$). This variability is negligible with respect to the mismatch effects observed on the chip.

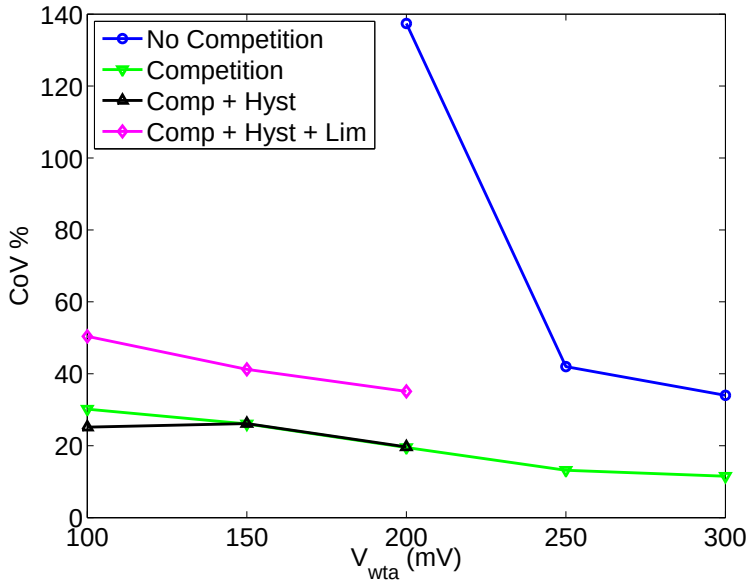


Figure 4.10: Evaluation of mismatch on the output of the I&F neuron circuits. The plot shows the coefficient of variation of the output frequency over the whole array for different settings of the competition as listed in the legend. The mismatch over the output frequency decreases when all the biasing transistors are in parallel (when the competition is enabled), and when the input current to the neurons is limited.

(see Fig. 3.2) that source the input current into the membrane capacitor C_{mem} of the I&F circuit (see Fig. 4.7), and from the I&F circuit itself.

Any mismatch on the output I&F neuron frequency is relevant, since it influences the settling time of self–inhibition. Eventually, mismatch between inhibitory synapses causes variation in the dynamics of the IOR.

4.3.2 Input synapse characterization

In this section the DPI synapse is characterized in the framework of its use in the SAC, as the input block to the current–mode WTA circuit. In these experiments I set the biases $V_{wta} = 200$mV, $V_{inh} = V_{dd}$, and $V_{lim} = 2.85$V, and explored the effect of different bias setting for the input synapse. As WTA competition is instantaneous and continuous in time, I looked for synaptic parameters that rendered the WTA input currents as smooth as possible, minimizing the peaks corresponding to each input spike. If the currents have large peaks, the result of the competition depends on the relative timing of the input spikes, independent of their input frequency, and the network oscillates. Fig. 4.11 shows some examples of time traces of the node V_{net}, which gives a measure of the input current to the WTA cell, for different values of the synapse time constant and weight. A long time constant allows the integration of the input spikes at the synaptic level (i.e. on the capacitor C_{syn}), rendering the output current smoother. Fig. 4.12 shows the mean I_{net} versus the input frequency, when the synapse is stimulated with spike trains of constant frequency. The mean current is estimated from the transfer function of the diode–connected transistor M_{net} of Fig. 3.2: $I_{net} = I_0 e^{\frac{kV_{net}}{U_T}}$, with the realistic assumption of transistor operated in subthreshold and saturated (Liu et al., 2002). The shaded areas in the figure show the standard deviation of the

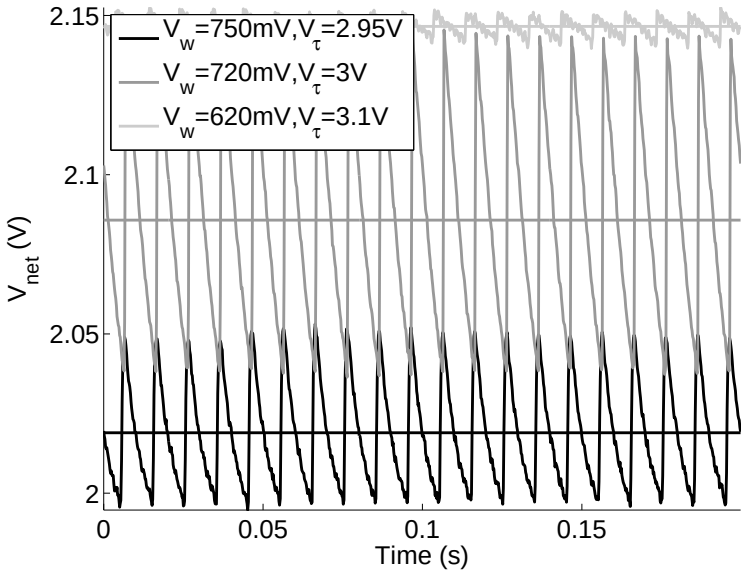


Figure 4.11: Input excitatory synapse: traces in time of the voltage V_{net}, and their mean, when stimulating the input synapse with a spike train of constant frequency at 100Hz, for three different bias settings. The traces correspond to the highlighted points of Fig. 4.12.

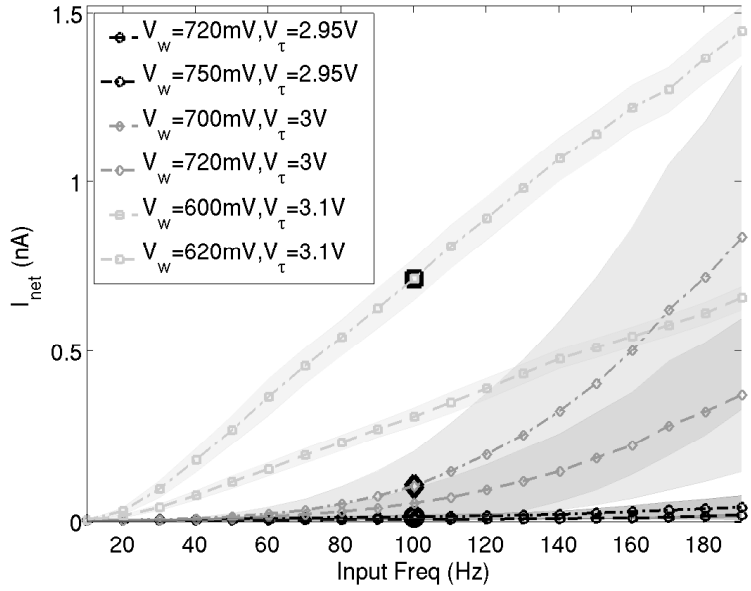


Figure 4.12: Input current to the WTA cell for different time constant and weight settings: mean and standard deviation (shaded areas) of the input current of the WTA cell versus the input frequency, when stimulated with trains at constant frequency. For long time constants the input current is proportional to the input frequency for a range comprising very low input frequencies, and the slope is sufficiently steep to discriminate the inputs. In addition, increasing the time constant decreases the standard deviation of the current, leading to smaller oscillations around the mean value when the input spike arrives. The highlighted points correspond to the traces of Fig. 4.11.

current[2]: for long time constants (i.e. V_τ close to the supply voltage, V_{dd}) the output current is smoother and the mean current is directly proportional to the input frequency for a wide range of input frequencies.

Another issue that needs to be taken into account when choosing the bias settings for the synapse is the magnitude of the current I_{net} it generates: four currents converge in the input node of the WTA cell: the excitatory synapse current; the inhibitory synapse current; the positive feedback current from the WTA; and the current from the lateral transistors, belonging to the resistive grid. All these input currents sum linearly, obeying to the Kirchhoff current law. If the total current sourced into the WTA circuit is too high, the node voltage increases. The drain of the M_{hyst} transistor of Fig. 3.2 is also connected to this node; if it increases above $(V_{dd} - 4U_T)$, the transistor goes out of saturation (Liu et al., 2002) and its current sharply drops, reducing the positive feedback effect. Therefore I choose a sensible bias setting for the input synapse, corresponding to small amplitudes of the synaptic excitatory current: $V_{w0} = 620$mV, $V_{thr} = 3.1$V, and $V_\tau = 3.1$V (corresponding to the light gray curves of Fig. 4.12).

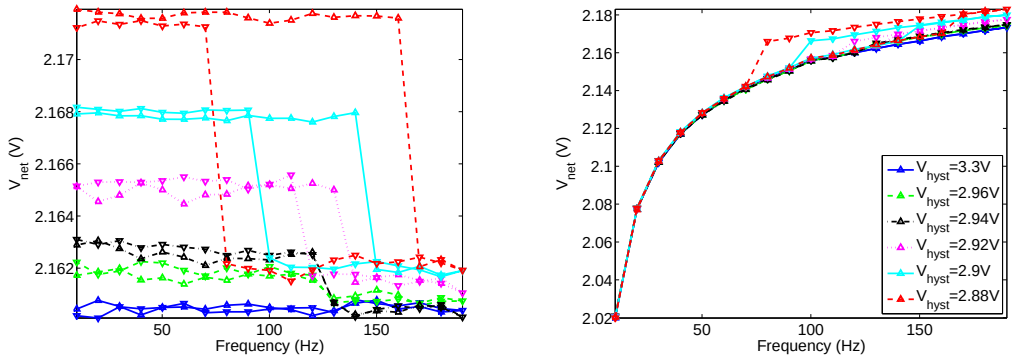
4.3.3 Hysteresis characterization

With the experiments described in this section, I characterized the effect of the positive feedback current on competition between two pixels of the array, quantifying the effect of the bias, V_{hyst}, that controls the magnitude of the hysteretic current.

Pixel vs Pixel

To uncover the effect of the hysteretic current, pixel 1 was stimulated with a regular train of spikes at a constant frequency of 100Hz, and pixel 2 with input frequencies increasing from 10Hz to 200Hz, and back, with steps of 10Hz. Without hysteretic current ($V_{hyst} = V_{dd}$), the WTA should switch from pixel 1 to pixel 2, when pixel 2 receives an input frequency slightly higher than 100Hz. The WTA should then switch back to pixel 1 when the input frequency of pixel 2 decreases back to a value slightly lower than 100Hz. When the hysteretic current is enabled, the winning pixel receives an extra current in the input node. This is equivalent to having an input spike train to the winner of frequency at $(100 + \Delta f)$Hz, where Δf depends on the bias V_{hyst}. In such a case, pixel 2 can win the competition only if its input frequency increases above $(100 + \Delta f)$Hz. As pixel 2 is selected by the WTA, the hysteretic current is removed from pixel 1 and is injected in the input of pixel 2. For pixel 2 to be deselected, its input frequency has to decrease below $(100 - \Delta f)$Hz. Fig. 4.13 shows the voltage of the WTA input node V_{net} recorded from the two stimulated pixels in these conditions. Its value depends exponentially on the input frequency and on the hysteretic current. Pixel 1 receives a train at constant frequency, therefore the contribution of the input synapse is constant throughout the experiment; when it is selected by the WTA, a constant value that depends on V_{hyst} is added to the trace. In the trace of pixel 2 one can observe the superposition of the synaptic response to the raising input frequency with the hysteretic current, when the pixel wins the competition for saliency. The different traces correspond to different values of V_{hyst}; as expected, when $V_{hyst} = V_{dd}$ (and the hysteretic current is off) the curves for increasing and for decreasing the input frequency are equal, while for smaller values of V_{hyst} the hysteretic effect becomes evident. Fig. 4.14 shows the V_{net} traces

[2]The standard deviation of the mean over time of the current is calculated with the standard procedure for error propagation: if the derived measure is $f(x, y, ...)$, where $x, y, ...$ are the direct measures, its standard deviation is $\sigma = \frac{\partial f}{\partial x}\sigma_x + \frac{\partial f}{\partial y}\sigma_y + ...$



(a) (b)

Figure 4.13: Hysteresis measured by observing the input node of the WTA: V_{net} mean value (over time) of pixel 1 (a) and pixel 2 (b), versus the input frequency of pixel 2. Pixel 1 is stimulated with a spike train of constant frequency at 100Hz, pixel 2 with a spike train that every 4s increases with a 10Hz step. (a) Pixel 1 receives always the same mean input current and its input node results constant for increasing the input frequency of the other pixel. When the hysteretic current is turned on, the input node of the pixel receives an extra current while winning; this current increases exponentially for decreasing the gate voltage of the transistor M_{hyst}, as shown by the different level of V_{net} (during the winning phase) obtained for decreasing values of V_{hyst}. The different level of V_{net} when the pixel is loosing depends on a leakage current in the feedback branch, that increases with the bias V_{hyst}. The hysteretic effect given by the activation of the feedback current is revealed in the different path obtained for increasing and decreasing the input frequency of the second pixel. (b) In the input node of pixel 2 there is the superposition of the hysteretic current with the response of the synapse to different frequencies.

of the two pixels superimposed, with and without hysteretic feedback. The plots show that the inputs to the WTA pixels cross for an input frequency of pixel 2 different from 100Hz, determined by the mismatch in the input synapses. In the baseline condition (Fig. 4.14(a)), the crossover frequency is the same for increasing and decreasing the input frequency of pixel 2. When hysteretic feedback is enabled (Fig. 4.14(b)), the crossover frequency moves towards higher values for increasing the input frequency of pixel 2, and towards lower values for decreasing it. The effect observed in the input node by measuring V_{net} is reflected in the overall behavior of the network. Fig. 4.15 shows the hysteretic curve of the output activity of the array, when the input frequency of pixel 2 increases linearly from 10Hz to 200Hz and back with a resolution of 1Hz, as shown in Fig. 4.15(a).

The hysteretic curves of Fig. 4.15(b) are obtained by plotting the center of mass (CoM) of the network activity versus the input frequency of pixel 2, where CoM is defined as

$$CoM = \frac{\sum_i \theta_i f_i}{\sum_i f_i}, \tag{4.1}$$

where θ_i is the address of the i^{th} pixel and f_i is its output frequency. When the input frequency of pixel 2 is low, pixel 1 wins the competition and the center of mass of the network corresponds to pixel 1; as the input to pixel 2 increases, the input currents to the two pixels start being comparable. Depending on the resolution of the WTA network, both pixels are selected as winners for a given range of frequencies, and the center of mass is in between the two; when the input frequency to pixel 2 increases enough above the input frequency to pixel 1, the network selects pixel 2 as winner, and the center of mass shifts to pixel 2. When the hysteretic current is disabled ($V_{hyst} = V_{dd}$), the *transition frequency* of the WTA, defined as the input frequency at which the WTA changes from one winning pixel to another, depends on the mismatch between the input currents generated by the excitatory synapses

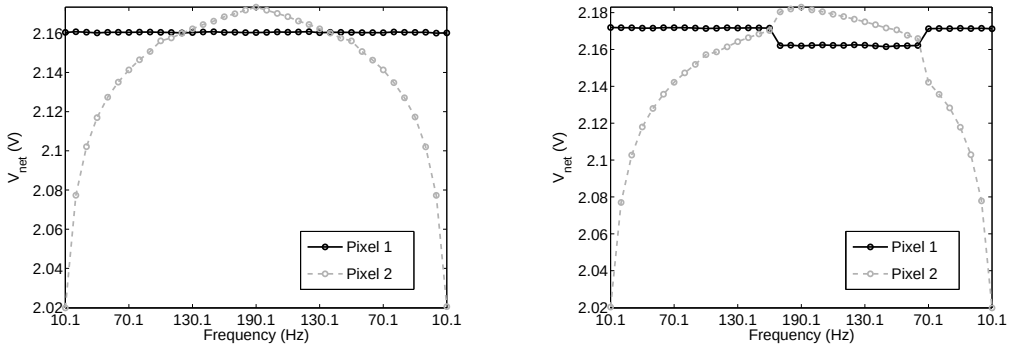


Figure 4.14: Hysteresis measured by observing the input node of the WTA: mean voltage of the input node of the WTA for the two stimulated pixels, versus the input frequency of pixel 2. (a) Baseline activity, when the feedback current is off: the input currents of the two pixels do not cross exactly when the input frequencies are equal, but when the input frequency of pixel 2 is about 120Hz, because of the mismatch between the input synapses. This results in a shift of the transition point of the WTA circuit around 125Hz, as shown in Fig. 4.15. (b) $V_{hyst} = 2.88$V: for low values of the input frequency of pixel 2, the input current to pixel 1 is higher than in the baseline, therefore the cross point of the two currents shifts towards higher input frequencies. When the network changes winner, the hysteretic current is removed from the previously winning pixel and is sourced into the current winner. For the same principle the crossing point for decreasing the input frequency shifts towards lower frequencies.

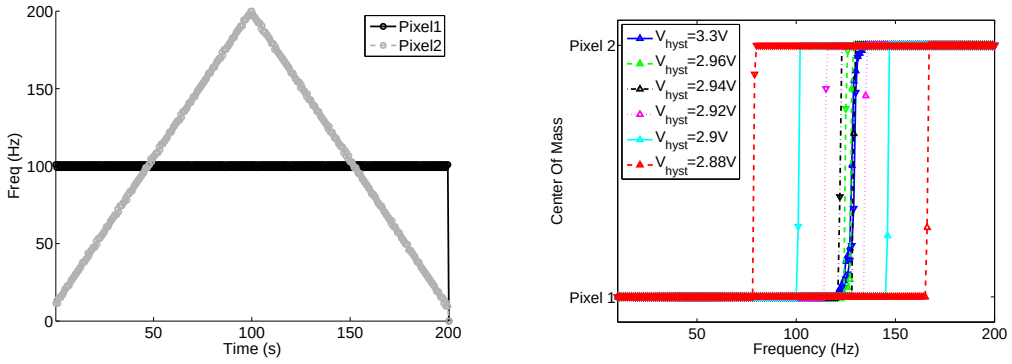


Figure 4.15: Hysteresis measured by observing the output activity of the I&F neuron: (a) Instantaneous input frequency of the spike train sent to pixel 1, and to pixel 2.The last point has an error due to the discretization of the spike train for the evaluation of the instantaneous firing rate. (b) Center of mass of the chip's activity versus the input frequency of pixel 2 for different amplitudes of the hysteretic current. The values of the bias V_{hyst} are the same used in the previous experiment. When the feedback current is disabled ($V_{hyst} = V_{dd}$) the network should change winner at 100Hz, where the two input frequencies cross, in these measures the network switches around 125Hz, because of the mismatch between the two input synapses, as shown in Fig. 4.14. The resolution of the WTA circuit is about 15Hz: within this range it cannot resolve between the two inputs, and both pixels are active. The activation of the feedback current increases the resolution of the network up to 1Hz, and produces the hysteretic phenomenon: when the hysteretic current is strongest ($V_{hyst} = 2.88$V), pixel 2 wins when its input frequency increases of $\Delta f \simeq$50Hz above the transition frequency, then in the way back, it looses when its input frequency decreases of the same Δf below the transition frequency. The value of Δf depends on the hysteretic current.

of the two pixels. In this experiment the transition point is around 125Hz, and is the same for increasing and decreasing values of the input frequency of pixel 2. When the hysteretic current is enabled, the transition frequency shifts toward higher values for increasing values of input frequency, and toward lower frequencies for decreasing values of the input frequency. The width of the hysteretic curve increases with increasing hysteretic current amplitude. Another phenomenon uncovered by this experiment is the increase of the WTA resolution: when the hysteretic current is active, the transition between one winner and the other is sharper than in the baseline, as there is typically only one point in the center of mass of the activity that corresponds to the activation of both pixels.

4.3.4 Lateral Excitation

Here I characterize the effect of the lateral facilitating connections between neighboring pixels of the WTA, implemented by means of a diffusive grid shown in Fig. 3.2, and described in Sec. 4.2.3.

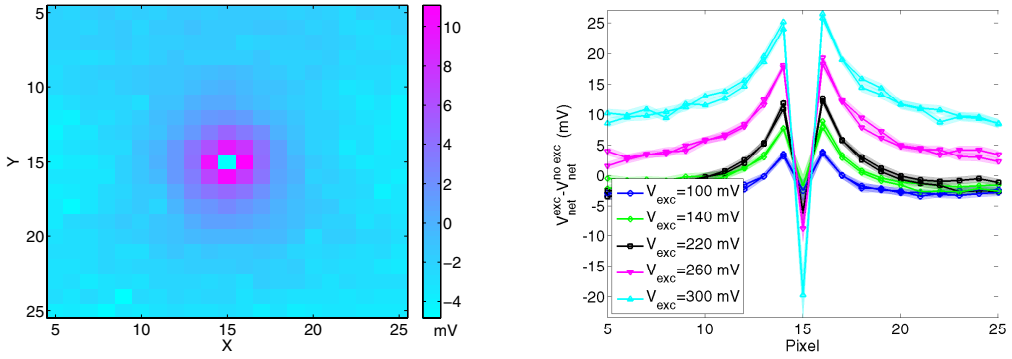
With the first experiment I characterize the extent of the lateral diffusion of the current with different values of the bias V_{exc}, then I examine the functional role of lateral facilitation, showing that a region of activity has a competitive advantage over a single pixel. The last experiment of this section shows that spatial smoothing of the input currents helps in reducing the effect of mismatch in the input synapses.

Lateral diffusion of the input current

To measure the spatial extent of the input current for different values of the parameter V_{exc}, I stimulated the central pixel (15,15), and measured the voltage V_{net} of the surrounding pixels. Fig. 4.16 shows the difference between the mean of V_{net}, obtained for a fixed value of V_{exc}, and its baseline, obtained when the lateral excitation is disabled ($V_{exc} = Gnd$). Fig. 4.16(a) shows the data of all recorded pixels, for $V_{exc} = 200$mV. The first neighbors of the stimulated pixel are brighter, then the amount of current received by more distant pixels decreases sharply. Fig. 4.16(b) shows mean and standard deviation of the same data, recorded from the pixels belonging to the same row and column as the central pixel (15,15), for a subset of values of V_{exc}. The value of the central pixel is negative, since the lateral spread of its input current decreases the net current to the pixel itself. For high values of lateral excitation the current spreads equally to the limit of the array, and the input loses any information about the location of the stimulated pixel.

Pixel vs Blob

Here I show the functional role of lateral facilitation in the WTA competition. Similar to the experiment described in Sec. 4.3.3, pixel 1 is stimulated with a constant frequency, while pixels belonging to an area of 3×3 centered around pixel 2, from now on referred to as *blob*, are stimulated with spike trains with frequencies ranging from 10Hz to 200Hz with steps of 1Hz and back. Fig. 4.17 shows the center of mass of the activity of the chip shifting between the single pixel and the blob for different values of the bias V_{exc}, and without the hysteretic current. When the lateral excitation is enabled, the pixels belonging to the same blob cooperate and the transition frequency of the WTA selection shifts toward lower values, until the blob wins even when stimulated with a lower frequency than the pixel. Thanks to the lateral facilitation a contiguous region of activity (e.g. arising due to an object) has a competitive advantage over a single pixel of activity.



(a) (b)

Figure 4.16: Lateral excitation. Spatial impulse response of the WTA resistive grid (see Fig. 3.2). Pixel (15,15) is stimulated with a constant spike train at 100Hz, and V_{net} of each pixel is recorded. The difference between the response for $V_{exc} > 0$ and the response for $V_{exc} = Gnd$ is plotted. The negative peak in correspondence of pixel (15,15) shows that part of the input current of the pixel diffused to the neighboring pixels. (a) Example of the spatial response for $V_{exc} = 200$mV, (b) Cross section with mean and standard deviation of the data recorded from the pixels belonging to the same row and column as the central pixel (15,15), for different values of the bias V_{exc}.

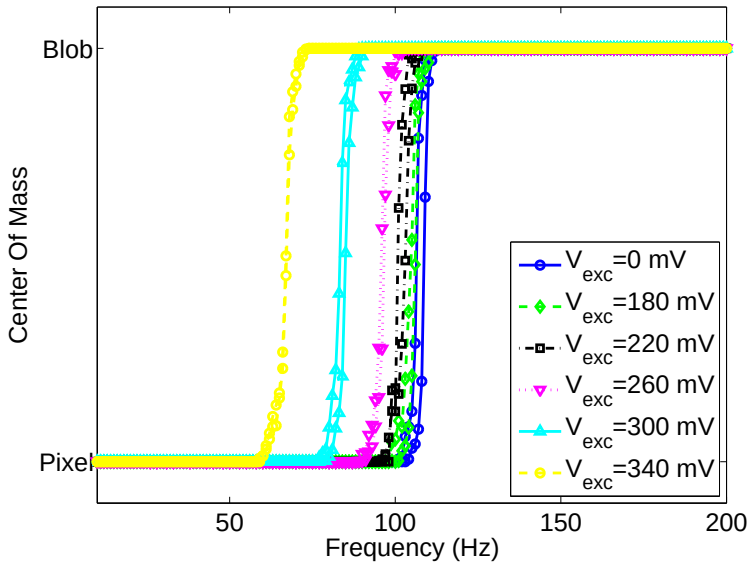
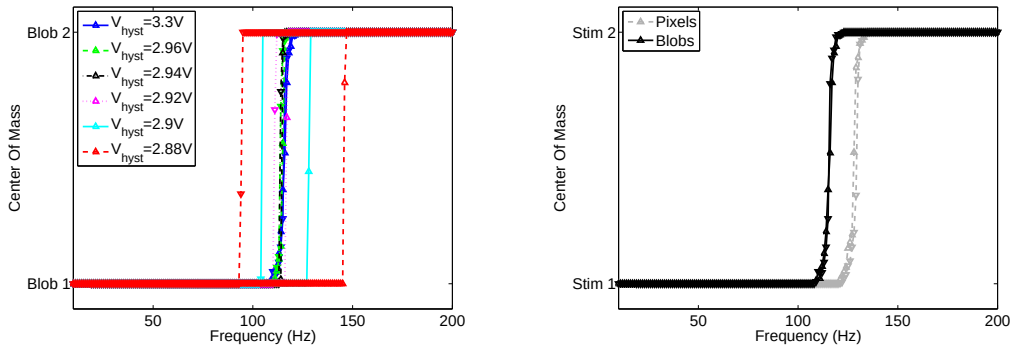


Figure 4.17: Functional role of lateral excitation: center of mass of the activity of the array, when stimulating a single pixel with a constant frequency and a blob with a frequency that increases and then decreases linearly, for different values of the bias V_{exc}. The activity of all of the pixels belonging to the blob is added together, and represented as a single pixel. When the lateral facilitation is enabled, the network transition frequency moves toward lower values. The hysteretic current (for $V_{hyst} = 2.9$V, not shown) has effect only for very small values of lateral excitation, since also the hysteretic current diffuses to the neighbors and becomes negligible.



(a) (b)

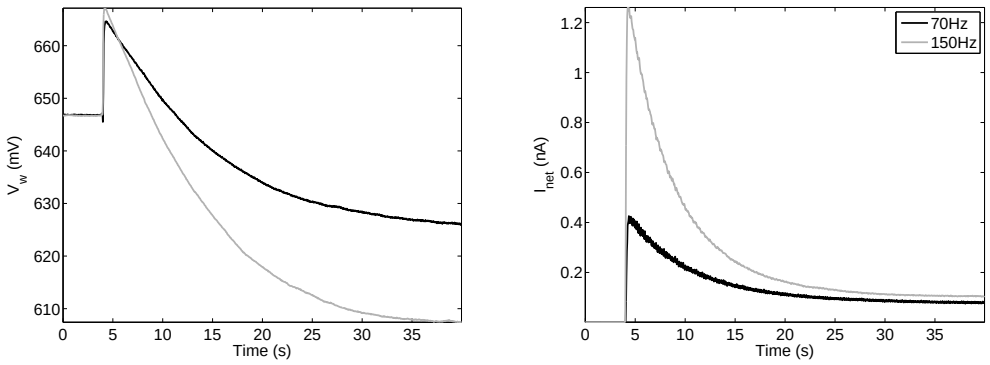
Figure 4.18: Hysteresis and Lateral Excitation with blobs: center of mass of the chip activity, where "blob 1" and "blob 2" correspond respectively to the region of activity around pixel 1 and pixel 2. (a) The center of the hysteretic curves is shifted toward 100Hz, corresponding to a reduced mismatch on the input currents of the WTA circuit. For high values of the hysteretic current the asymmetry of the hysteretic curve is due to the mismatch on the feedback branch. (b) Baseline activity without hysteresis, when either single pixels or blobs are stimulated, for a direct comparison: The transition frequency of the WTA is shifted about 15Hz.

Blob vs Blob

This last experiment was performed to unveil an additional role of lateral facilitation, spatial smoothing, which reduces the effect of mismatch between pixels. As pointed out in Fig. 4.15, the transition frequency of the network does not correspond exactly to the input frequency of pixel 1 (100Hz), but is shifted toward a higher frequency because of mismatch between the input currents generated by the two synapses (see also Fig. 4.14). A way to reduce this effect is by using lateral excitation to smooth the input in space. Fig. 4.18 shows the same experiment as in Fig. 4.15, stimulating the two blobs centered around pixel 1 and 2. The transition frequency of the WTA selection shifts toward the input frequency of pixel 1 (100Hz), showing a reduction of the effect of mismatch between the input currents to the WTA pixels. Fig. 4.18(b) shows the baseline behavior when the hysteretic current is disabled: when stimulating two blobs, the network selects the new winner when it is stimulated at a frequency closer to 100Hz than when stimulating only two single pixels. Fig. 4.18(a) shows the hysteretic curves obtained in this experiment; for the same value of V_{hyst} the width of the hysteretic curve is smaller, since the hysteretic current spreads also to the neighbors: The slight asymmetry in the hysteretic curves with respect to the baseline transition is probably due to mismatch between the hysteretic currents.

4.3.5 Short–term depression

Here I characterize the effect of short–term depression (STD) of the synaptic weight, implemented on the input excitatory synapses. The STD mechanism, described in detail in Sec. 2.4.1, adapts the weight of a synapse with an increasing number of input spikes. In terms of bottom–up attention, the saliency of stimuli whose attributes never change is decreased, and stimuli with one or more varying attributes become more salient, implementing a form of visual adaptation (McDermott et al., 2006). In the first experiment I measured the effect of the short–term depression circuit on the weight of the synapse (V_w in Fig. 4.6), and on the output current of the input synapse in the test pixel (0,31). The second experiment quantified the effect of STD on the competition between two pixels, when the input



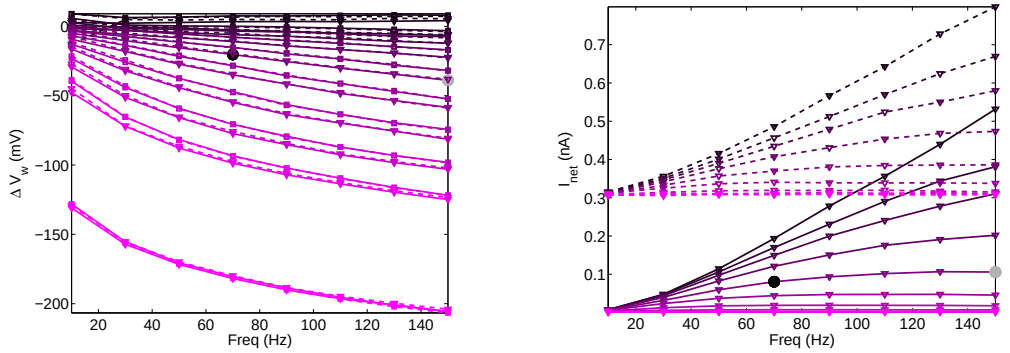
(a) (b)

Figure 4.19: Short–term depression: plot of the synaptic weight V_w (a), and of the synapse output current I_{net} (b), when the synapse is stimulated with two different input frequencies. The initial synaptic weight is set to $V_{w0} = 620$mV, the strength of the depression is $V_{wstd} = 240$mV, and the hysteretic current is turned off. The measured initial value of the synaptic weight is about 645mV, the discrepancy with respect to the value externally set is most likely due to offsets on the read–out buffers. At the onset of the stimulation ($t = 5$s), there is an initial positive peak, due to parasitic effects in the STD circuit. Subsequently the weight decreases toward a steady state value that depends on the input frequency.

frequency of the loser undergoes a step change.

Single Pixel

To characterize the effect of the short–term depression circuit on the variables directly involved, I measured the steady state value of the synaptic weight and of the synapse output current I_{net}, when stimulating the synapse of the test pixel (0,31) for different values of the STD strength V_{wstd}, and of the input frequency f_{in}. For each run of the experiment the synapse was reset by turning off the short–term depression circuit and waiting for the weight to recover, then the new value of V_{wstd} was set and the synapse stimulated with a constant frequency. Fig. 4.19(a) and 4.19(b) show examples of the V_w and I_{net} traces respectively, for a fixed set of biases and for two different input frequencies, corresponding to the highlighted dots in Fig. 4.20(a) and 4.20(b). Fig. 4.20 summarizes the results obtained by sweeping V_{wstd} and f_{in}, for two different values of the initial synaptic weight V_{w0}, both with and without the hysteretic current. As expected, the steady state value of the weight is independent of the hysteretic current and of the initial synaptic weight; it decreases for increasing input frequencies, and for increasing values of the bias V_{wstd}. The result of the synaptic weight depression is the decrease of the output current; I evaluated this current by measuring the total input current I_{net} to the corresponding WTA node, when the lateral connections and the inhibitory feedback were turned off. This measure reflects the synaptic contribution and the hysteretic current. Fig. 4.20(b) shows the output current of the synapse when the initial weight is set to $V_{w0} = 620$mV; for increasing frequencies of the input train the steady state amplitude of the current decreases. The effect increases for stronger synaptic depression, set by the bias V_{wstd}. Following the same protocol, I quantified the effect of a step change in the input frequency on the output current: first the synapse was stimulated with a fixed frequency, then, after the synaptic weight reached the steady state, the input frequency was increased; the experiment was repeated for 5 increasing values of the frequency step, starting from two different frequencies, for two different values of V_{w0}, and for different values of the bias V_{wstd}. Fig. 4.21 shows the change in the output current versus the



(a) (b)

Figure 4.20: Short–term depression, weight variation (a) and steady state value of the synaptic current (b) as function of the input frequency. Each curve corresponds to a different value of the bias V_{wstd}: the black trace corresponds to $V_{wstd} = 0$mV, the brightest trace to $V_{wstd} = 400$mV, the intermediate curves to values ranging from $V_{wstd} = 180$mV to 320mV with 20mV increments. The markers distinguish the initial value of the synaptic weight: $\triangledown$ for $V_{w0} = 620$mV, $\square$ for $V_{w0} = 680$mV. The dashed lines correspond to the experiments with the hysteretic current: (a) as expected the hysteretic current has no effect on the synaptic weight, but only on the output current; (b) when the input current is small the hysteretic current sums linearly and the I_{net} curves are translated of about 300nA. (a) The weight variation depends on both the input frequency, and the bias V_{wstd}, but not on the initial synaptic weight V_{w0}. (b) The steady state of the current depends linearly on the input frequency and the slope is determined by the steady state value of the synaptic weight.

amplitude of the frequency step. As expected from the analytical derivation in Sec. 2.4.1, the synapse responds to changes with a variation of its mean output current proportional to the amplitude of the input change.

Pixel vs Pixel

With this experiment I characterize the effect that STD has on the competition between two stimuli. Pixel 1 initially received a spike train at a constant frequency of 100Hz and pixel 2 received a train at a lower frequency (50Hz); after 60s the input frequency to pixel 2 increased step–wise for different amplitudes of the step, ranging from $\Delta F = 5$Hz to $\Delta F = 75$Hz. Fig. 4.22 shows the raster plots of the two pixels (pixel 1 in blue), for different values of the final input frequency to pixel 2, in four representative cases.

In Fig. 4.22(a) the pixel receiving the highest absolute frequency wins, until there is a step change in the input frequency of pixel 2. In this case pixel 2 wins transiently, even if its absolute frequency is still lower than the frequency of pixel 1; when the absolute frequency of pixel 2 is above the absolute frequency of pixel 1, pixel 2 wins, but as its weight depresses, the output currents of the two pixels become comparable, and both pixels win.

In Fig. 4.22(b) the V_{wstd} bias is stronger. The steady state of the synaptic weight is lower for pixel 1 than for pixel 2, because the first receives a higher input frequency: pixel 2 starts to win even before the input frequency step. This behavior is even more pronounced when V_{wstd} is increased further (see Fig. 4.22(c)). Here the weight of pixel 1 depresses after just a few seconds and pixel 2 wins; when the frequency of pixel 2 increases above 95Hz, the weight of pixel 2 depresses to a value comparable to the weight of pixel 1, and both the pixels win. For higher values of the absolute input frequency of pixel 2 the value of the synaptic weight decreases further, and pixel 1 wins the competition. Fig. 4.22(d) shows the response of the WTA for a moderate value of depression ($V_{wstd} = 220$mV), and with the

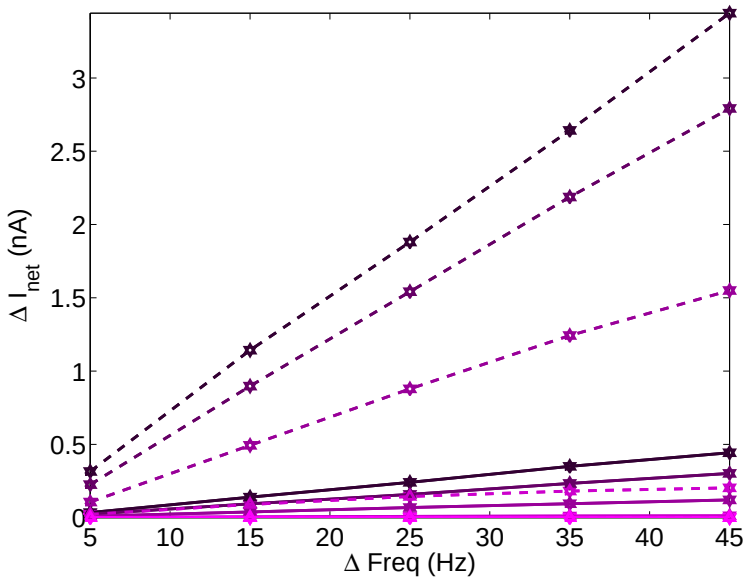


Figure 4.21: Short–term depression, dependence of the current output on the variation of the input frequency. The synapse of pixel (0,31) is stimulated with frequency steps of increasing amplitude, the correspondent step in the output current ΔI_{net} is plotted versus the amplitude of the frequency steps $\Delta Freq$. The solid lines correspond to the initial weight $V_{w0} = 620$mV, the dashed lines to $V_{w0} = 680$mV; the brightness of the curves increases with the bias V_{wstd}, the darker corresponds to $V_{wstd} = 0$mV, the brighter to $V_{wstd} = 400$mV, and the intermediate values range from 200mV to 280mV with increments of 40mV. The markers $\triangle$ and $\triangledown$ correspond to different initial values of the input frequency, respectively 50Hz and 70 Hz. The curves with different markers superimpose in all the conditions, confirming that the current variation depends on the value of the frequency step.

hysteretic current enabled. In this case, pixel 2 wins only when the step sets its absolute input above the winner's frequency. Hysteresis clearly separates the selection behavior in two regions: either pixel 1 wins always; or pixel 2 wins after the frequency step, and the network stabilizes again. Fig. 4.23(a) and 4.23(b) show the center of mass of the network activity versus the duration of the activity of pixel 2 , for increasing depression strength (V_{wstd} ranging from 200 to 300mV, with 10mV increments), and for each ΔF, with for initial stimulation frequencies of pixel 2 of 50Hz and 70Hz. The points at zero duration and centered around pixel 1 show that pixel 2 was never selected; the points in the left half of the space correspond to the case of Fig. 4.22(a), where pixel 2 wins after the input frequency step, and the duration is related to the amplitude of the step. There are clusters around 60s; in this case pixel 2 is active after the frequency change until the end of the experiment, however the center of mass of the network indicates that both pixels 1 and 2 are active, and the network does not select a winner. The last group corresponds to the experiments of Fig. 4.22(b) and 4.22(c), when pixel 2 starts to win before the step, because the weight of pixel 1 is depressed to a very low value. The comparison between the data corresponding to the two experiments with different initial input frequencies confirms that the output current of the depressing synapse, and therefore the result of the competition, depends on the amplitude of the step change, ΔF, not on the final absolute value of the input frequency.

4.3.6 Inhibition Of Return

In this section I characterize the dynamics of the inhibition of return (IOR) mechanism. The IOR mechanism is implemented with local inhibitory synapses (see Sec. 4.2.5, Fig. 4.8); the

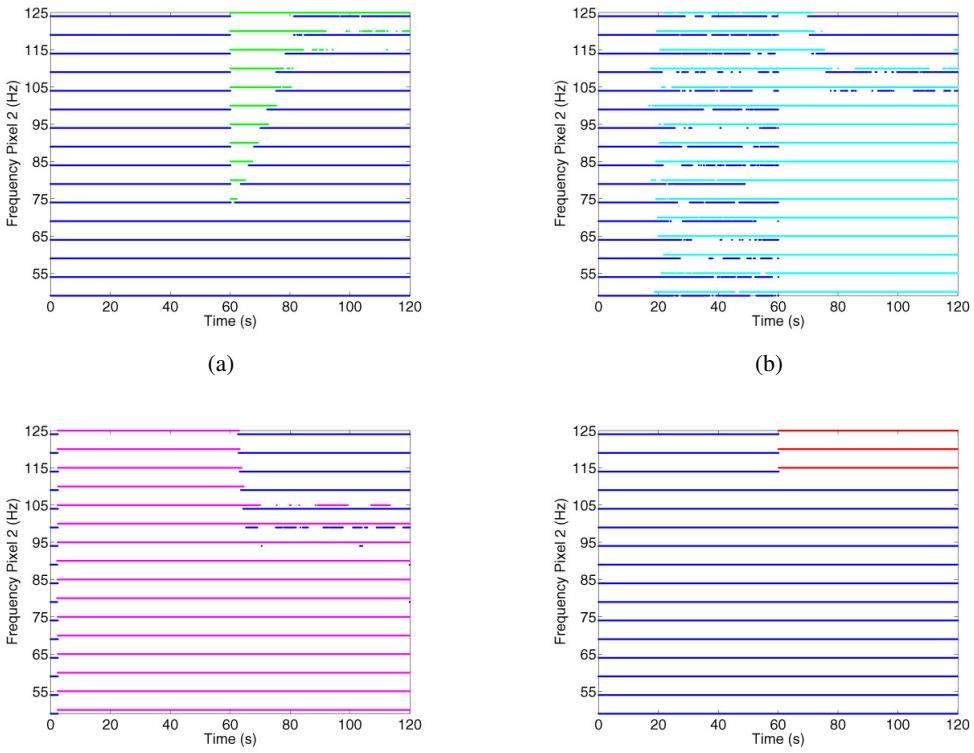
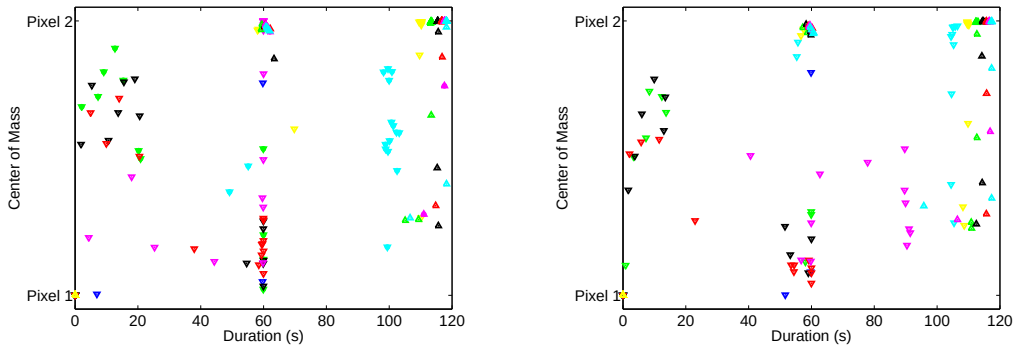


Figure 4.22: Short–term depression, response of the SAC when two pixels are stimulated at different constant frequencies and the input to the losing pixel undergoes a step change ΔF of increasing amplitudes. Each figure shows raster plots of the activity of the stimulated pixels for different values of the final absolute frequency of pixel 2, obtained for ΔF ranging from $\Delta F = 5$Hz to $\Delta F = 75$Hz. Pixel 1, plotted in blue, is always stimulated at 100Hz, pixel 2, plotted in different colors, is stimulated first at 50Hz, then after 60s its frequency increases step–wise to 50Hz+ΔF. The plots in (a), (b), and (c) are obtained for initial synaptic weight $V_{w0} = 620$mV, and without hysteretic current, for three different values of the depression strength: (a) $V_{wstd} = 200$mV, pixel 2 wins when the frequency step is sufficient high, and is active only during the transition, before being depressed in turn. (b) $V_{wstd} = 240$mV, and (c) $V_{wstd} = 290$mV, pixel 1 depresses after few seconds and pixel 2 wins even before the frequency step; pixel 1 wins again when also pixel 2 is depressed. In (d) the hysteretic current is turned on, and the WTA switches only for high ΔF ($V_{wstd} = 240$mV).

winning pixel activates the output I&F neuron, whose spikes are integrated by the inhibitory synapse. This subtracts current from the input node of the WTA cell, until its net input current is lower than the input currents of the other pixels in the array. Another pixel can then be selected by the WTA network. The efficacy and time constant of the inhibitory synapse, set by V_{winh}, V_{thrinh}, and $V_{\tau inh}$ respectively, control the dynamics of self–inhibition. They determine how many spikes are required for the WTA to deselect the current winner and select another pixel, and also set the time constant of the inhibitory current.

Single Pixel

The dynamics of IOR can be characterized by measuring the duration of the activation of the I&F neuron and the duration of its suppression. The instrumented pixel (0,31) was stimulated with a spike train of constant frequency of 100Hz, and the IOR dynamics were evaluated for different values of V_{winh} and $V_{\tau inh}$. The duration of activation and suppression of the I&F neuron are plotted in Fig. 4.24, and in Fig. 4.25 where the hysteretic current is



(a) (b)

Figure 4.23: Short–term depression, center of mass of the array versus the duration of the activity of pixel 2. Pixel 1 is stimulated with a spike train at 100Hz, Pixel 2 is first stimulated at 50Hz (a), or 70Hz (b), and after 60s its frequency increases of ΔF. Each color–marker combination correspond to one experiment with a given bias setting, for different values of ΔF. Points in the left half space of the plots corresponds to transient activity of pixel 2 (see Fig. 4.22(a)), that wins for a duration proportional to ΔF. Points in the right half space correspond to the behavior shown in Fig. 4.22(b) and 4.22(c). The region at the center of the plots is a transition. The highlighted dots correspond to the raster plots in Fig. 4.22.

enabled ($V_{hyst} = 2.85$V). Fig. 4.26 shows the time course of the internal variables monitored from the instrumented pixel during the experiments.

The lower trace of the plots in Fig. 4.26 shows the membrane potential V_{mem} of the output I&F neuron (see Fig. 4.7). Depending on the bias settings of the inhibitory synapse, one or more spikes are sufficient to inhibit the current winner.

The upper trace corresponds to the input node voltage of the WTA V_{net} (see Fig. 3.2), and shows the decrease of the input current to the current winner due to the inhibitory current.

The middle trace shows the gate voltage of the output transistor of the inhibitory synapse (M_{ior} in Fig. 4.8). A linear decrease of this voltage results in an exponentially decaying current subtracted from the input node of the corresponding WTA cell; depending on the time constant and on the weight of the synapse, the output inhibitory current pulse will last from a few milliseconds up to two seconds.

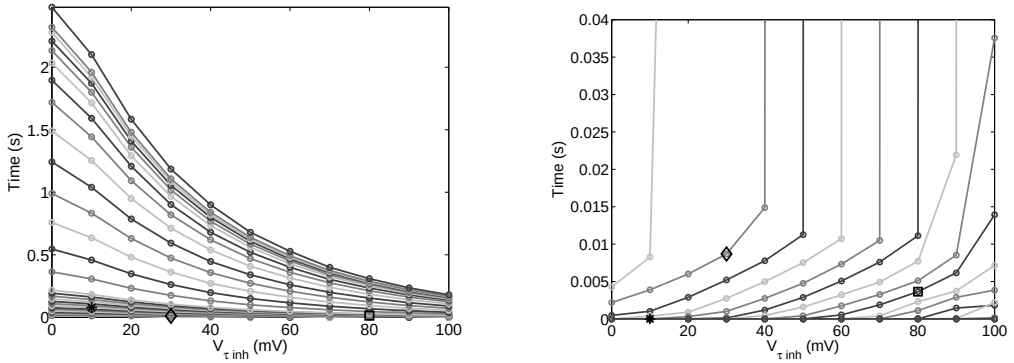
4.4 Conclusions

In this chapter I characterized the properties of the circuits used in the SAC with control experiments, and studied their behavior as a function of the main tunable parameters.

I showed how the output current of the input excitatory synapse is linear with stimulus saliency. Specifically, I determined the synaptic parameters that render the synapse output current suitable for the current–mode WTA.

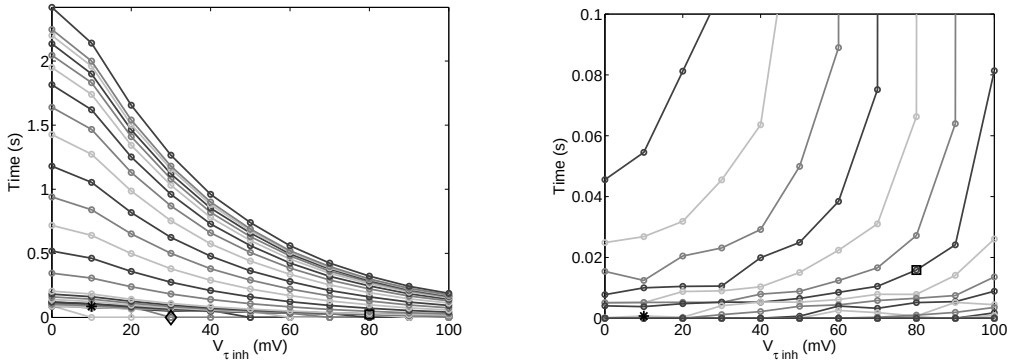
As for any physical system, including biological ones, the SAC shows intrinsic inhomogeneities in its constituent elements, resulting in reduced precision and variation in the dynamics of self–inhibition. In Sec. 4.3.4 I demonstrated the use of lateral excitation in the hysteretic WTA circuit for the reduction of mismatch; this experiment is especially relevant to confirm once again the importance of cooperative computation to reduce the effects of low precision in physical systems.

In Sec. 4.3.3 I characterized the effect of hysteresis in increasing the resolution and stability of the WTA, as predicted by the circuit analysis in Sec. 3.3.



(a) (b)

Figure 4.24: Inhibition of return: Quantification of the effect of the inhibitory synapse parameters, the weight V_{winh} and the time constant $V_{\tau inh}$, monitoring the spiking activity of the instrumented pixel, stimulated with a spike train of constant frequency at 100Hz. (b) Mean duration of the spiking activity, and (a) the mean duration of the silent period, versus $V_{\tau inh}$; each curve corresponds to a different value of the synaptic weight. Increasing $V_{\tau inh}$ decreases the synaptic time constant, and the duration of inhibition, while the duration of activation of the neuron increases; increasing V_{winh} increases the weight of the synapse and the duration of inhibition, on the contrary the duration of neuronal activation decreases, as expected from the synaptic circuit transfer function. The markers $*$, $\Diamond$, and $\Box$ correspond to the traces in time shown respectively in Fig. 4.26(a), 4.26(b), and 4.26(c).



(a) (b)

Figure 4.25: Inhibition of return with hysteresis, same experiment of Fig. 4.24, for ($V_{hyst} = 2.85$V). The duration of the suppression period (a) is unaffected, the duration of the activation period (b) increases, given the additional positive current fed into the input node of the WTA. The markers correspond to the bias settings of the correspondent markers in Fig. 4.24

In Sec. 4.3.5, I showed the relevance of the newly introduced short–term depression circuit in the input synapses. STD makes the chip sensitive to the sudden appearance of a stimulus, and to changing stimuli.

Finally in Sec. 4.3.6 I characterized the dynamics of the IOR mechanism.

The experiments described in this chapter show the functional effects of various parts of the chip, both in the context of circuit stability and performance , and in the context of attentional selection, for controlled inputs. They are relevant for isolating the contributions of the different parts of the system, which typically interact in a highly non–linear fashion. This approach mimics the early psychophysics strategy of studying the visual system with visual search arrays and artificial stimuli, unveiling the basic principles and characteristics of the system in a controlled environment. Validation of the SAC as a tool for attentional selection, when all its parts are interacting and dealing with complex situations, requires an

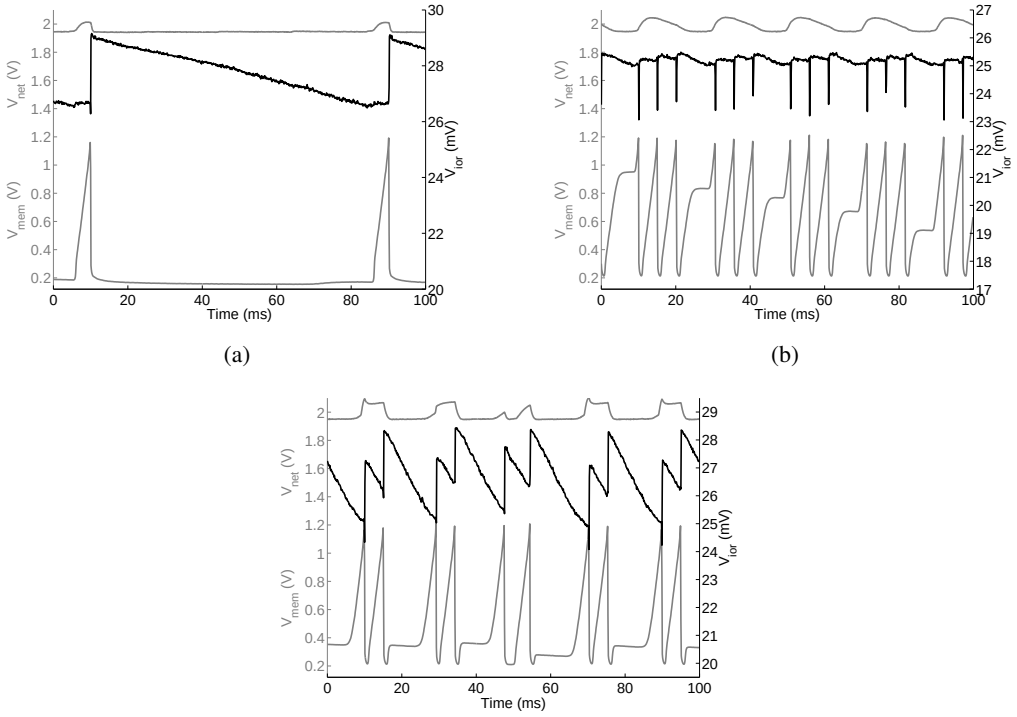


Figure 4.26: Inhibition of return: typical traces of the internal variables V_{net} (top trace), V_{mem} (bottom trace), and V_{ior} (black middle trace), recorded from the instrumented pixel, for three combinations of V_{winh} and $V_{\tau inh}$. The input current from the excitatory synapse causes the increase of the current in the input node of the WTA cell (V_{net}), that wins the competition for saliency; the I&F fires action potentials (V_{mem}). The inhibitory synapse integrates the spikes: V_{ior} increases and a current is subtracted from the input node (V_{net}), causing the deselection of the pixel as winner; the neuron stops firing. The cycle starts again as soon as the inhibitory synapse turns off. (a) $V_{winh} = 2.42$V, $V_{\tau inh} = 10$mV, (b) $V_{winh} = 2.58$V, $V_{\tau inh} = 30$mV,(c) $V_{winh} = 2.44$V, $V_{\tau inh} = 80$mV.

additional experiment, incorporating natural stimuli and observation of the overall behavior when the SAC interacts in real–time with the real world, as described in Chap. 5.

Chapter 5

A multi–chip selective attention system

5.1 Introduction

In the previous chapters I described the circuits implemented on the SAC and characterized their functional behavior. In this chapter I describe the behavior of the SAC when it is part of a saliency–map based selective attention multi–chip system, comprising a neuromorphic vision sensor and an actuator that orients the sensor towards the salient stimuli selected by the SAC.

The Focus of attention (FOA) scan paths produced by the SAC are strongly dependent on its input saliency map. To characterize the scan paths produced by the SAC in a controlled way, I generated synthetic saliency maps in software (see Sec. 5.2). I then implemented a two–chip system for selective attention, by connecting the SAC directly to an AER silicon transient imager (Lichtsteiner et al., 2004), which produces a proto–saliency map based on the information of local luminance changes in time (see Sec. 5.3). I used this system to evaluate the scan paths generated with real stimuli, and the behavior of the system when it interacts with the real world.

Following is a review of the state–of–the–art for both software and hardware implementations of saliency map based models of selective attention.

5.1.1 State–of–the–art implementations of saliency–map models

The Itti and Koch (2000) saliency–map based model of selective attention described in Sec. 1.2 has been implemented in software by (Itti et al., 1998). The *iLab Neuromorphic Vision C++ Toolkit (iNVT)*, released by Itti's group at the University of Southern California, implements all stages of the model from front–end data acquisition and feature map computation to the saliency map scanning that produces the final FOA scan path. The iNVT software implementation was used to validate the Itti and Koch (2000) model in different applications (Itti and Koch, 2000), and with comparison to human psychophysics (Itti, 2005a,b). An alternative implementation of the algorithm has been developed by Dirk Walther (Koch and Walther, 2006) at Caltech, who released it as the *Matlab SaliencyToolbox*. The Toolbox comprises the core functionality of the Itti and Koch (2000) saliency–map model but has the advantage of being independent of the platform used. Nevertheless, the number of feature channels supported for computation of the saliency map is limited compared with the iNVT implementation. Due to its slower performance, it is recommended only for computing saliency maps of static images.

Software implementations are important and useful for characterizing possible computational models, and for validating the simulated theories. They can also be used for

applications that require off–line image processing, such as data compression (Itti, 2004).

In real world applications such as robotic vision, video surveillance, etc., where the main requirements are real time processing and low power consumption, software simulations on desktops are inadequate so dedicated approaches have been proposed.

Itti's group at iLab developed a real–time implementation of the bottom–up saliency model on a 16–CPU Beowulf cluster (Itti, 2002), that runs the iNVT software; with optimized dedicated libraries it can achieve 205Mbps transfer rates, that leads to a speed of 30 frames per second (for 320×240 image size). This approach guarantees high performance and the flexibility typical of software simulations, allowing the implementation of the full saliency map model, extended with object recognition, or any other type of algorithm for vision. The disadvantage of this approach are size and power requirements.

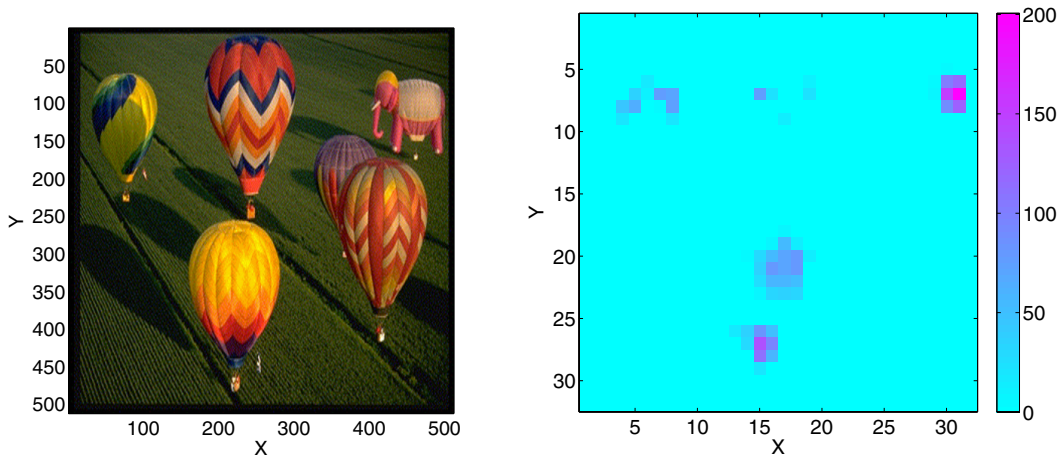
Park et al. (2003) implemented a saliency map based model for selective attention using a CCD camera connected to an IBM PC and mounted on a DC motor. The images acquired with the camera are transferred to the PC, which selects the next location to be attended; a control algorithm implemented on a DSP (digital signal processor) drives the DC motor for the camera foveation.

A hybrid approach, based on a dedicated hardware platform but still using software modules, has been adopted by Ouerhani et al. (2002): data from a digital CMOS imager are processed by a cluster of 4 Single Instruction Parallel Data (SIMD) machines, the ProtoEye. Each SIMD has highly parallel and low power hardware. The operations of this architecture are controlled by a 4MHz general purpose microcontroller. A simplified version of the saliency map model was implemented on this system, adapting the algorithm to the hardware. The system can process 14 frames per second, at a resolution of 64×64 pixels; it has lower speed and resolution than the Beowulf implementation, but requires less power and is still suitable for many applications.

Another approach is to implement the model on dedicated analog VLSI hardware, that directly maps the required algorithm on silicon. On one hand this approach has less flexibility, since the hardware cannot be fully reprogrammed to implement different algorithms, even though parameter tuning can change the operating conditions of the networks. On the other hand it has the great advantage of real–time and low power operation. Within the neuromorphic engineering approach, as underlined also in Chap. 4, there are two main streams: one strategy is to implement, on a single chip, the sensors and post–processing that implements a simplified selective attention model using a very small set of features, such as local stimulus intensity (Brajovic and Kanade, 1998; Morris et al., 1996) or temporal derivative of contrast (Horiuchi et al., 1997; Indiveri, 1999), to compute the saliency map. The second approach, adopted in this work, is to separate the sensory acquisition and processing stages, realizing multi–chip hierarchical systems that have the advantage of higher flexibility (Indiveri, 2000a). The SAC can be used to implement the last decision stage of the saliency map based model, by hard WTA competition and an IOR mechanism, needed to produce the FOA scan path. However it can also be tuned to operate as a local WTA, without IOR, to implement the required local competition and normalization in the feature map component of the saliency map model. Multiple instances of the SAC tuned for different features can be used to build a hierarchical selective attention multi–chip system.

5.2 SAC response to synthetic saliency maps

In this section I evaluate the behavior of the SAC when stimulated with saliency maps generated in software with the algorithm proposed by Itti and Koch (2000), and compare



(a) (b)

Figure 5.1: SaliencyToolbox: (a) Input image used for the following experiments (part of the image database of the SaliencyToolbox). (b) Saliency map relative to the input image, obtained from the SaliencyToolbox, with the default parameters.

the properties of SAC–generated scan paths to the scan paths obtained by the software implementation. Specifically, I show the effects of different IOR configuration settings on both the scan paths generated by the software algorithm and by the SAC. The following experiments exploit the SAC as a tool to explore the parametric space of the implemented model in real time.

5.2.1 Methods

I used the Matlab SaliencyToolbox for generating the saliency map from input images. Specifically the saliency maps were computed from *color*, *intensity* and *orientation* (at 0,45,90,135 degrees) feature maps, weighted equally and summed to generate the saliency map.

The saliency maps created with the SaliencyToolbox were transformed into an appropriate input for the SAC: for each pixel, a constant spike train was produced whose frequency was proportional to the saliency value of the pixel itself. To have an input range within the linearity region of the input synapses, I mapped the saliency values to an interval between 0Hz and 200Hz. As in the experiments of Chap. 2 and 4, spike trains were generated with the Matlab SpikeToolbox and sent to the SAC via the PCI–AER board, which was also used for monitoring the output activity of the chip (e.g. see Fig. 5.6).

5.2.2 Results

The benchmark image I used is shown in Fig. 5.1(a) (and is also used as a standard benchmark in the SaliencyToolbox). Fig. 5.1(b) shows the corresponding saliency map computed by the SaliencyToolbox with its default parameters. The toolbox also generates the focus of attention scan path. The default settings of the toolbox create an inhibition region for the IOR with the shape of the object selected. For a more direct comparison with the SAC, which lacks any top–down information regarding the concept of objects, I changed the inhibition region to be a disc centered around the FOA. Fig. 5.2 shows the resulting scan path for two different dimensions of the inhibition region: if the size is small (Fig. 5.2(a)),

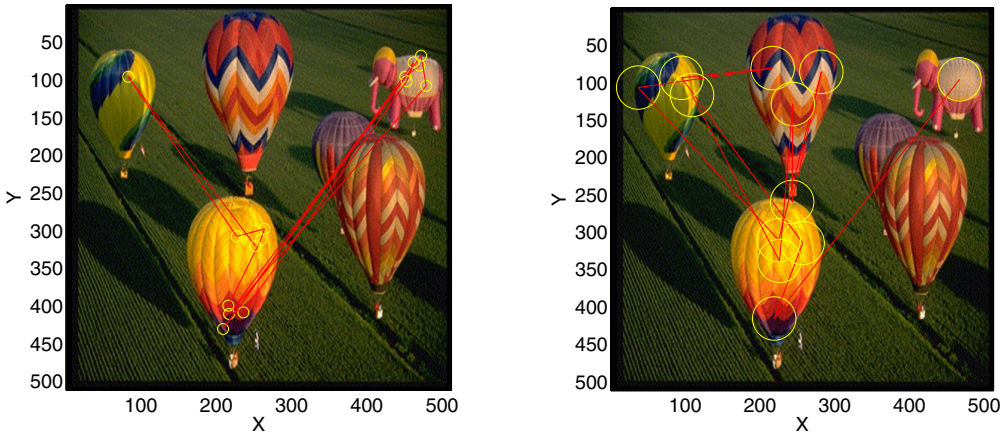


Figure 5.2: Focus of attention scan path generated by the SaliencyToolbox: The yellow circles are centered around each fixation point belonging to the FOA scan path, the red lines connect consecutive fixations. The radius of the yellow circles shows the size of the inhibition area. (a) When the inhibition area is small, many points belonging to the same object are selected, and some less salient objects are discarded. (b) A larger inhibition area prevents the network from selecting points belonging to the highest saliency region, allowing the selection of stimuli with lower saliency.

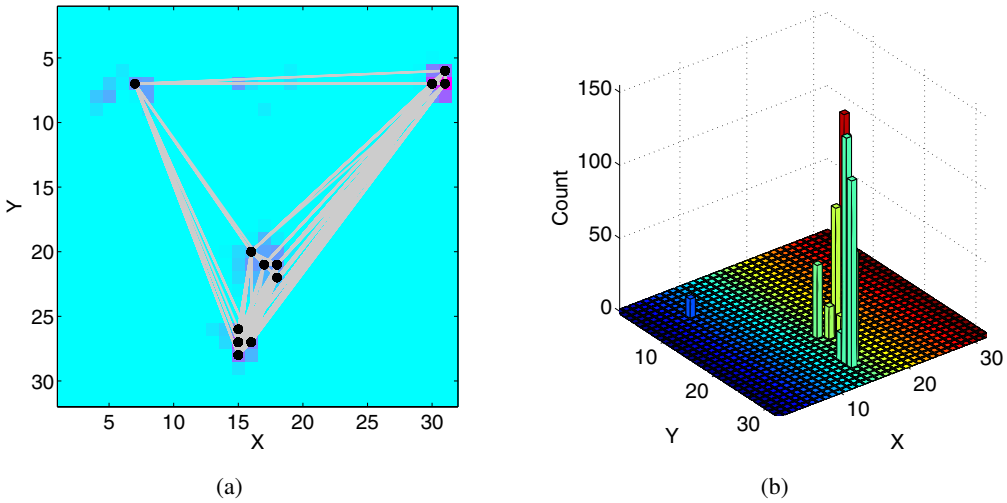


Figure 5.3: Focus of attention scan path generated by the SAC: (a) FOA scan path superimposed on the saliency map. The black dots show the fixation points, the grey lines connect consecutive fixations. (b) Histogram of the visited points in the saliency map, the chip selects more often regions with higher saliency, qualitatively reproducing the scan path observed in Fig. 5.2(b).

the algorithm chooses many points belonging to the same few objects with high saliency and avoids less salient targets. If the size is bigger (Fig. 5.2(b)) less salient locations are visited. A similar behavior is also observed in the scan path generated by the SAC, when stimulated with the same saliency map: Fig. 5.3(a) and 5.4(a) show the focus of attention scan path generated by the SAC superimposed on the saliency map for two different sets of parameters, which differ in the time constant of inhibition. In the SAC, the space constant of the lateral excitation also contributes to the lateral spread of the inhibitory current, but the inhibition region decays exponentially within few pixels and depends on the amplitude of the current itself, as shown in Sec. 3.3 and 4.3.4; therefore it is not possible to set ar-

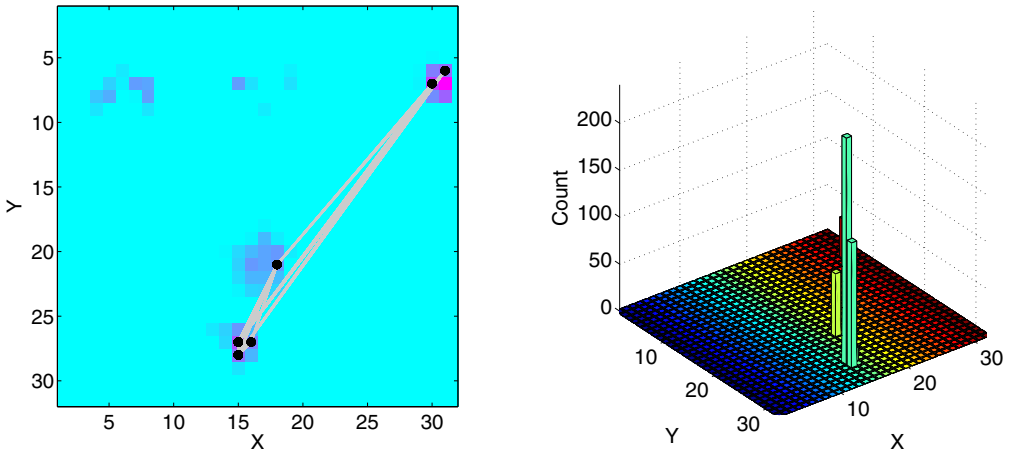


Figure 5.4: Focus of attention scan path generated by the SAC: the IOR time constant is faster, and the SAC selects a smaller set of pixels, qualitatively reproducing the behavior of the SaliencyToolbox for small inhibition regions, shown in Fig. 5.2(a).

bitrary sizes for the inhibition region. The number of different pixels selected in the scan path, including lower saliency regions, increases if the time constant of IOR is longer: the inhibitory current forces the input to the inhibited pixel to a low value, the network does not select previous winners for a long time, and less salient pixels can be selected. The SAC with global competition, lateral excitation, hysteresis, and IOR, qualitatively reproduces the scan paths generated by the SaliencyToolbox, but selecting regions of higher saliency more often.

The experiments performed here show that we can use the SAC as a tool for implementing the final decision stage of the saliency map model, and use it to explore the parametric space, in real time.

5.3 Two–chip system response properties

In this section I describe the behavior of the AER–based multi–chip system for modeling selective attention I have developed.

The system for the first time combines a neuromorphic sensor chip, a post–processing chip that performs a non–trivial computation over the sensory data, and an actuator driven by the result of the sensory data post–processing.

The AER sensor, a silicon transient imager, referred to as the *silicon retina* in the following, produces a proto–saliency map based on the local changes in contrast.

The silicon retina was designed by Lichtsteiner and Delbrück (2005) at the Institute of Neuroinformatics. It generates asynchronous events corresponding to temporal changes in the logarithm of local image intensity. As $\frac{d}{dt}\log I = \frac{dI/dt}{I}$, where I is the pixel illumination, the retina output encodes temporal changes in contrast rather than absolute illumination differences; this property allows the retina to adapt to the global illumination level, responding to 20% contrast over a dynamic range spanning over 5 decades. Each pixel of the retina responds to both positive and negative variations in contrast, transmitted as ON and OFF events respectively. In the specific implementation of the two–chip system proposed in this thesis, ON/OFF information is discarded and the events sent to the SAC indicate a variation

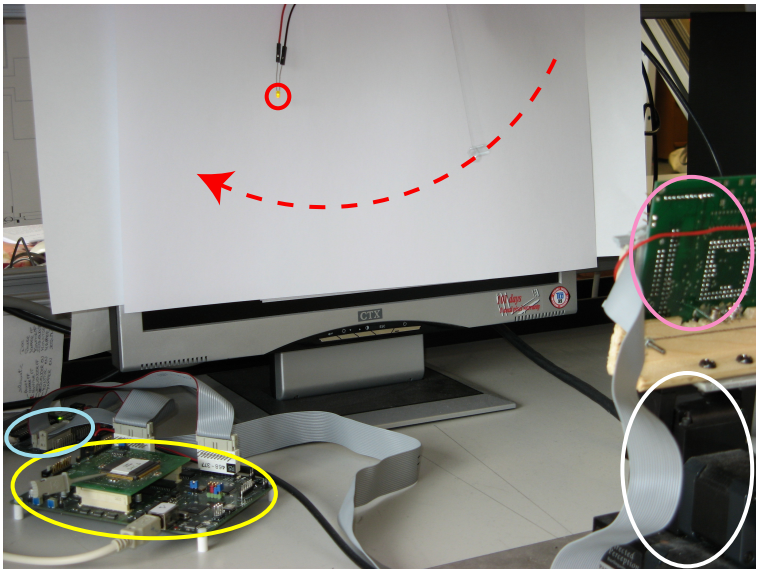


Figure 5.5: Selective attention multi–chip system: the AER silicon retina (pink ellipse) is mounted on a Pan–Tilt Unit (white ellipse), it is connected to the SAC (yellow ellipse) via the PCI–AER board (light blue ellipse), that connects both chips to a Linux desktop. In front of the retina, over a white background, there are a flickering LED (red circle) and an oscillating nut (red dashed arrow), used as stimuli for the experiment described in Sec. 5.3.2.

of contrast, irrespective of its polarity.

The address events of the retina are sent directly to the SAC to determine where the focus of attention has to be deployed. The actuator, a Pan–Tilt Unit (PTU), orients the silicon retina towards the FOA location selected by the SAC.

5.3.1 Methods

Fig. 5.5 shows a picture of the experimental setup; it comprises the silicon retina mounted on a Pan–Tilt Unit, the SAC, and the PCI–AER board connected to a Linux PC. The realization of such a complex system was allowed by custom software and hardware infrastructures, developed by different groups for AER–based neuromorphic chips. In Appendix C I describe in detail how I merged the parts developed within different frameworks, and show all of the components of the multi–chips system (see Fig. C.1). In this experiment, activity of the AER chips was recorded for off–line analysis and used on–line to control the PTU. The retina has an array of 64×64 pixels; the spikes produced by the retina were routed to the SAC via the PCI–AER mapper functionality: a look–up table implemented a 4:1 mapping of the retina addresses onto the 32×32 addresses of the SAC. The next location of the focus of attention was determined by a software algorithm, based on the spiking activity of the SAC, acquired via the PCI–AER monitoring functionality. The resulting coordinates of the FOA were transformed into a motor command, and sent to the Pan–Tilt Unit connected to the PC via a serial port, to shift the central pixels of the retina to the coordinates of the new focus of attention.

In the first experiments the FOA scan path was recorded while maintaining the retina fixed at its initial fixation point, reproducing what in psychophysics experiments is known as *covert* attention (see also Sec. 1.2).

In the last set of experiments the focus of attention was used to foveate the retina, reproducing *overt* attentional scan paths (Findlay, 2005).

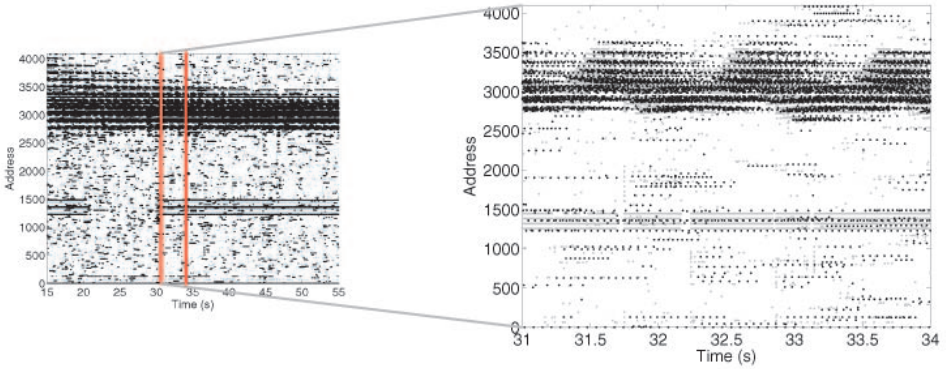


Figure 5.6: The raster plots show the activity of the SAC (black dots) superimposed to the activity of the retina (grey dots). The three lines with addresses ranging from about 1200 to 1500 correspond to the LED evoked activity, the addresses belonging to the higher band of events (ranging approximatively from 2800 to 3500) correspond to the nut swing. (a) The activity is plotted for 50 seconds, showing the LED turning off and then on again. (b) Zoom over three seconds after the LED turning on. These data correspond to the baseline experiment, where competition, as well as lateral excitation, hysteresis, and IOR, are turned off; the chip maps the input synaptic activity to the I&F neurons, that respond independently from the strength of the synaptic input.

5.3.2 Covert attention with Short–Term Depression

Covert attention is the term used in psychophysics experiments to indicate that the subject performs the task while maintaining his/her eyes fixed, typically looking at a central fixation point. In this section I describe covert attention experiments, where I measured the focus of attention (FOA) scan path generated by the SAC, while maintaining the retina focused on a fixed location.

As described in Sec. 4.3.5, STD in the input excitatory synapses renders the array sensitive to moving and changing stimuli, which are strong psychophysical attractors of attention (Itti, 2005a). This mechanism might appear to be redundant when the input to the SAC is provided from the transient silicon retina chip activity. Nevertheless, in many practical situations there are stimuli that produce activity at a constant frequency, therefore eliciting a strong activity in the transient retina chip (for example flickering lights); however such stimuli are stationary because their frequency does not change in time, therefore they should not be selected as strong attractors of attention.

One approach to overcome this problem is to remove activity corresponding to the stationary stimuli by using a band pass filter (Delbrück, 2006) tuned to their frequency. Short–Term Depression filters out inputs at high constant frequency. Using the STD properties of its input synapses, the SAC can be used to select the activity elicited by transient stimuli, whilst suppressing the activity of the stationary ones.

I reproduced a situation with one stationary stimulus and one transient stimulus, with the setup shown in Fig. 5.5: the LED flickering at a constant rate is used as a stationary stimulus, and the oscillating nut represents the "transient" stimulus.

The LED is driven by a function generator, at a frequency of 200Hz. During the experiment the LED is transiently turned off and, after a 10s pause, turned on again. Fig. 5.6 shows the raster plots of the retina (in grey) superimposed on the SAC activity (in black), for the baseline condition (i.e. without competition, lateral excitation, hysteresis, IOR, nor STD). Each active pixel of the retina elicits spiking activity in the corresponding I&F output neuron. Fig. 5.7 shows the SAC response with WTA competition, hysteresis, lateral excitation, and STD active, but without IOR ($V_{inh} = 3.3$V, $V_{wta} = 200$mV, $V_{hyst} = 2.9$V,

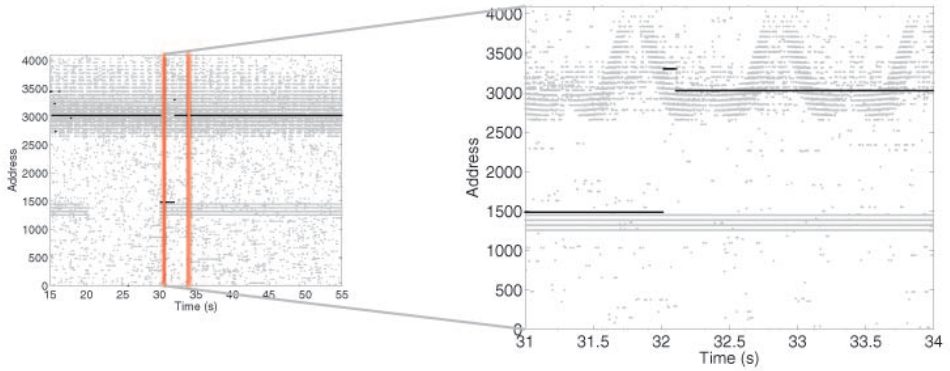


Figure 5.7: Raster plots of retina and SAC activity; the plots use the same convention as in Fig. 5.6. The response of the SAC is obtained for the chip operating with all its features turned on, except for IOR. The network selects only one pixel, and does not follow the moving stimulus; the LED is transiently more salient than the current winner, until the STD dims it.

$V_{exc} = 200$mV, $V_{wstd} = 280$mV). The SAC selects one pixel stimulated by the nut trajectory; the corresponding input frequency is not high enough to elicit a significant depression of the synapse weight, and hysteresis contributes to maintain the selection stable. When the LED is turned on (at $t = 30$s) the winner switches transiently, for as long as the weight of its input synapse is not suppressed by the STD. When STD decreases the weight of the pixels stimulated by the LED, the first selected winner is chosen again. Without the IOR mechanism, the SAC does not disengage from the selected pixel, unless a stronger input is applied. This configuration (with no IOR) is useful to evidence the transient effect of the appearance of a strong stationary stimulus.

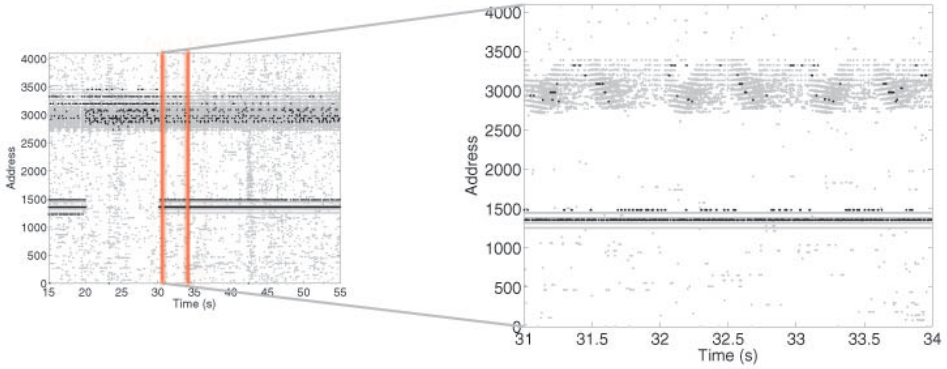
Fig. 5.8 shows the SAC response with the IOR mechanism also active ($V_{\tau inh} = 80$mV, $V_{thrinh} = 200$mV, $V_{winh} = 2.4$V). The three plots correspond to three STD conditions: absent ($V_{wstd} = 0$V), medium ($V_{wstd} = 280$mV), and strong ($V_{wstd} = 350$mV). Without STD, the WTA switches between the LED and the nut; when STD is too strong, the synaptic weight of pixels stimulated at high frequency is depressed, and pixels stimulated at low frequency win the competition. When an intermediate value of STD is applied, the LED is selected only for a short transient and the SAC follows the nut movement.

The behavior of the system in this experiment shows the high level effect of the implementation of a low–level feature such as STD: stationary inputs do not attract attention, but their transient activation is capable of capturing attention for a short period; therefore the system is still sensitive to the appearance of (potentially) behaviorally important stimuli.

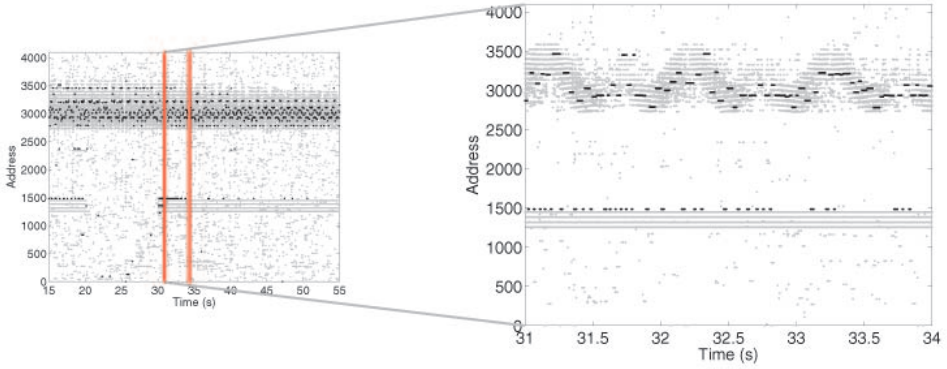
The superposition of retina and SAC activity shows a short latency of the SAC response, due to the dynamics of the interaction between IOR and hysteresis: when one pixel is selected, hysteresis keeps it active until the IOR inhibits the pixel and the WTA selects a new winner; in the meantime the stimulus has moved away, and the SAC follows it with a short latency. The latency could be corrected by modeling *predictive attentional saccades* (Itti, 2005a). For example, one could add to the address of the current winner an offset based on the mean speed of the nut.

5.3.3 Covert Attention with stimuli of different grey levels

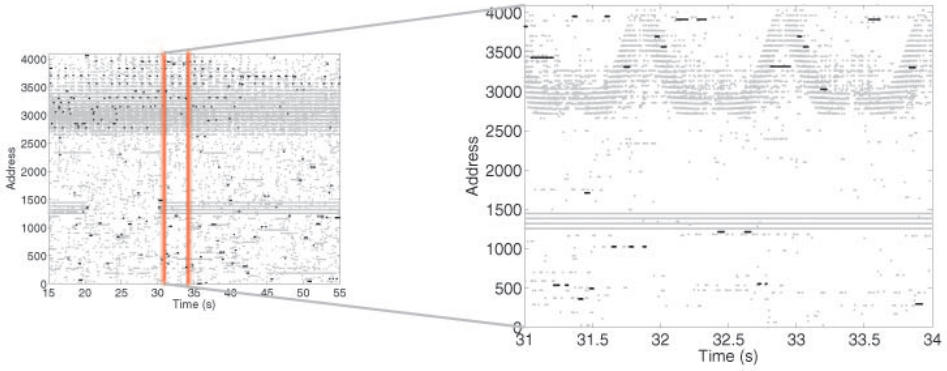
With these experiments I show a typical covert scan path generated in response to simple static images. Fig. 5.9 shows the two stimuli used. The images consisted of three circles over a white background: in the “baseline” condition all of the circles were colored in



(a)

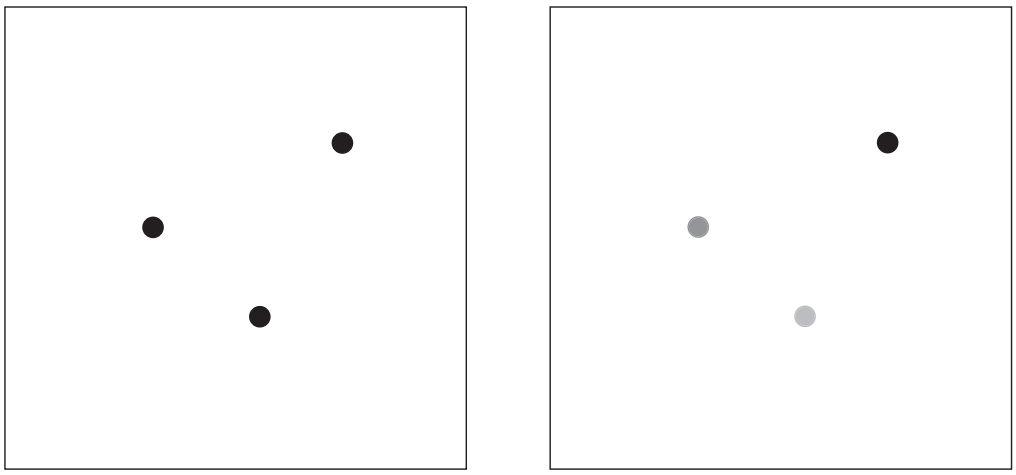


(b)



(c)

Figure 5.8: Raster plots of retina and SAC activity. The right column corresponds to time zoom of the left column, few swings of the nut can be observed in the top activity band. The plots are obtained with the same experiment settings of Fig. 5.7, with the addition of IOR and for three values of STD: (a) no STD ($V_{wstd} = 0$V), (b) medium STD ($V_{wstd} = 280$mV), (c) strong STD ($V_{wstd} = 350$mV).



(a) (b)

Figure 5.9: Stimulus for the covert and overt attention experiments. (a) Baseline: three black circles (grey level $gl = 1$); (b) Contrast: each circle has a different grey level, gl =0.3, $gl = 0.5$, and gl =1, corresponding to different levels of contrast over the white background ($gl = 0$).

black, in the "grey level" condition each one had a different grey level (black, grey, and light grey). Static images cannot elicit any activity in the retina, since by design it responds to temporal changes of contrast. Biology solved the issue of being sensitive to variations by continuously moving the retina with micro–saccades (Martinez-Conde et al., 2004, 2006). The same solution can be applied to the silicon system: the Pan–Tilt Unit is used to perform small displacements of the retina position in the x and y directions around the fixation point. The retina senses contrast variations at the edge of the circles.

I observed the FOA scan path generated by the SAC for two different settings of IOR: Fig. 5.10 shows the results for "slow" IOR, when the inhibition builds up with many spikes from the I&F neuron and the suppression time is long, thanks to a low synaptic weight and a slow time constant ($V_{winh} = 2.58$V, $V_{thrinh} = 200$mV, $V_{\tau inh} = 40$mV); Fig. 5.11 shows the results for "fast" IOR, when the inhibition builds up with a single spike of the I&F neuron (thanks to a strong synaptic weight), and recovers slowly, thanks both to a slow time constant and a strong weight ($V_{winh} = 2$V, $V_{thrinh} = 200$mV, $V_{\tau inh} = 80$mV).

The measure used to quantify the FOA scan path was the number of times each pixel was selected, normalized by the total number of shifts ("Activity(%)" in the plots).

As expected, in the "baseline" condition, shown in Fig. 5.10(a) and 5.11(a), all of the three circles were selected a similar number of times; the difference in the percentage between each circle depends mainly on the mismatch, both in the retina spike trains, and between the SAC pixels. In the "grey level" condition, shown in Fig. 5.10(b) and 5.11(b), the activity was proportional to the grey level of the circles: the black circle is selected more often than the grey circle, which in turn is selected more often than the light grey circle.

The two different IOR settings generate different scan path dynamics.

Fig. 5.10(c) and 5.10(d) show that for "slow" IOR the activity of each selected pixel is stable for few tens of milliseconds before being suppressed, resulting in about 20shifts/s of the FOA; similarly to the saliency map experiment of Sec. 5.2, only a single pixel corresponding to each circle is selected. The 3D plot of the scan path in Fig. 5.10(f) shows that the less active pixel is selected only after many FOA shifts.

Fig. 5.11(c) and 5.11(d) show that when the inhibitory synapse weight increases, each pixel is inhibited after a single spike of the corresponding I&F neuron; the inhibitory current

generated by the synapse, and subtracted from the WTA circuit input, has a high amplitude and slow decay. The effect on the WTA pixel is that the inhibitory current is active for a long time. With respect to the "slow" IOR settings, with the "fast" IOR settings, the relative duration of the inhibition of each pixel increases with respect to the duration of the pixel activity. More pixels are selected in the FOA scan path, similarly to the saliency map case of Fig. 5.4; the number of FOA shifts almost doubles, and in particular also the light grey stimulus is selected within the first few FOA shifts.

In conclusion, this set of experiments shows that the SAC is able to select salient stimuli, and that its output corresponds to the relative saliency of the different selected stimuli. For different configurations of the IOR settings the basic behavior does not change, but the dynamics of the scan path and the spiking activity of the winning pixels change. For "fast" IOR one or two spikes are emitted per winner, and the FOA shifts many times per second; for "slow" IOR each winning pixel emits spikes for a relatively long time window, but at the expenses of a decrease in the ratio between activity and suppression duration, that limits the number of pixels in the scan path. The differences in the spiking output of the SAC is important for the design of algorithms that read–out the SAC activity, as shown in Sec. 5.3.4.

5.3.4 Overt Attention with stimuli of different grey levels

The same stimulus of Fig. 5.9 was used for evaluating the performance of the system in the context of overt attention. In overt attention the FOA location guides the ocular movements, that orient the fovea of the retina towards the selected region of the input stimuli. In the two–chip system with the SAC and the silicon retina, the pixel selected by the SAC is used as target for the saccadic movement.

To model overt attention I developed a software algorithm, that acquires the activity of the SAC and translates it into commands for driving the Pan–Tilt Unit (PTU) to orient the retina accordingly.

As for the covert attention model, the silicon retina is moved performing micro–saccades around the current fixation point. Meanwhile the activity of the SAC is monitored for a temporal window of $\Delta t = 50$ms. The pixel that produced the maximum number of spikes during such acquisition window is chosen as the target for the next saccade; if there are more pixels who have the same maximum activity, the target is chosen randomly among them. The address of the selected pixel is translated in relative coordinates for the PTU and the motor command is sent to the PTU via the serial port. During the saccadic movement of the retina, the mapping from the retina to the SAC is temporarily disabled; this prevents all of the activity of the retina to be routed to the SAC during the saccadic retina movements, and corresponds to the implementation of a strong form of *saccadic suppression* (Ross et al., 2001). After the PTU movement, the PCI–AER mapping is re–enabled, the micro–saccades of the retina start again, and a new target for the next saccade can be selected.

Also this experiment was run with the "slow" and "fast" configurations of the IOR mechanisms. For the "fast" IOR settings, a single spike is emitted by the winner, and the winning pixel is inhibited for a long time. During the acquisition window many neurons are selected as winners by the SAC, each emitting a single spike. For this reason, in only about 10% of the cases there is a single pixel with maximum activity. While in about 90% of the cases there are multiple pixels whose activity corresponds to the maximum activity. For the "slow" IOR configuration, the winner elicits a burst of spikes (see Fig. 5.12). In such a case, during the acquisition window, the SAC selects very few pixels and in about 98% of

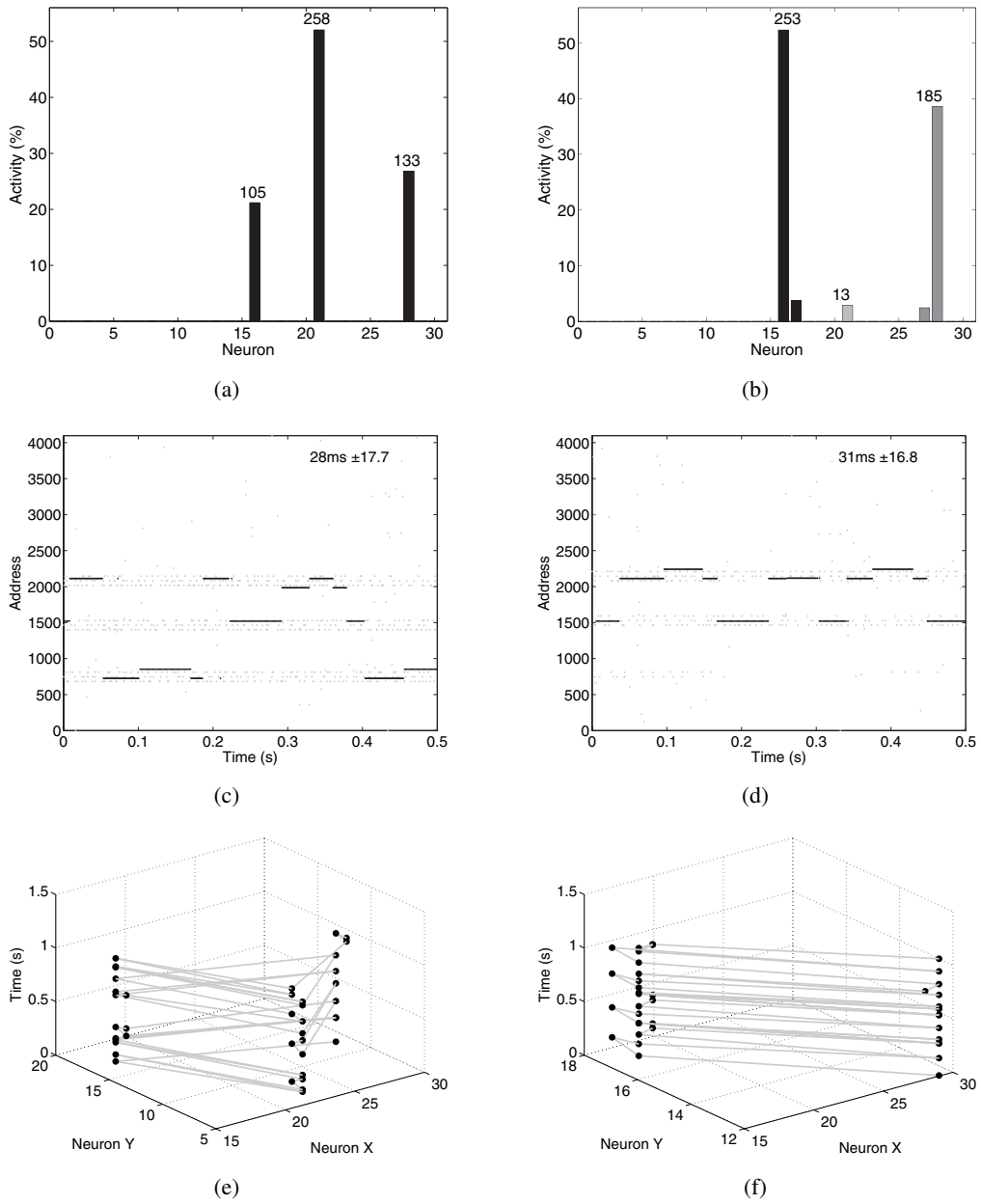


Figure 5.10: Covert attention scan path when the retina is stimulated with three circles. The left column plots correspond to the baseline case (all of the three circles black), the right column plots correspond to the three circles with different grey scale level. The top row shows the times each neuron address along the X axis of the SAC has been selected. The three circles can be distinguished on the basis of the addresses along the X axis, the grey level of each circle is reflected in the color of the bar relative to the corresponding neurons. The number over each bar shows the total number of times each of the three circles is selected. The middle row shows the raster plots, superimposing the retina (in grey) and the SAC (in black) activity; the activity corresponding to the three circles can be distinguished in the three bands of spikes, the lower corresponding to the middle circle in the histogram; the number on each plot shows mean duration of each pixel and its standard deviation. The lower row shows the pixels (black circles) selected over time (Z axis) for 1.5s. These experiments were performed with global competition, lateral excitation, and hysteresis enabled ($V_{inh} =$ 3.3V, $V_{exc} =$ 220mV, $V_{hyst} =$ 2.9V); the inhibitory synapse weight and time constant are small ($V_{winh} =$ 2.58V, $V_{\tau inh} =$ 40mV). In the "baseline" condition, all of the three circles are selected, the difference in the total number of times each pixel is selected depends on mismatch. When the three stimuli have different grey level, the activity of the retina pixels corresponding to the light grey decreases. The black circle is selected more often than the grey circle, which in turn is selected more often than the light grey circle. The plot of (f) shows that the pixels corresponding to the light grey circle are not selected for the first 1.5s of the experiment.

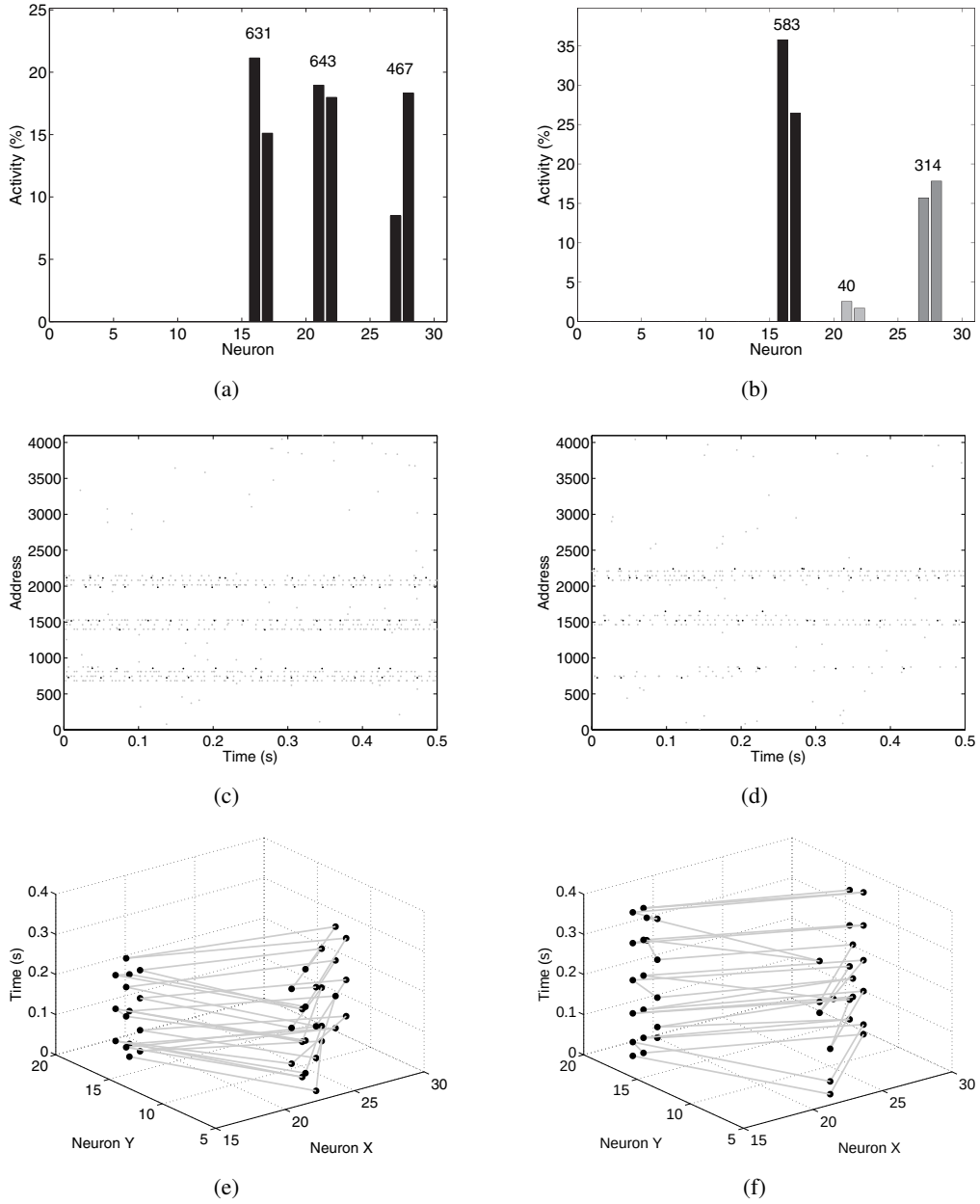


Figure 5.11: Covert attention scan path when the retina is stimulated with three circles. Same experiment as in Fig. 5.10, for a different configuration of the IOR dynamics: strong weight ($V_{winh} =$ 2V) and faster time constant ($V_{\tau inh} =$ 80mV). The effect is the same observed in Sec. 5.2: the number of different pixels selected has increased. The main difference with respect to the "slow" IOR settings of Fig. 5.10 is that the pixels corresponding to the light grey circle are selected within the first few FOA shifts (f), while in Fig. 5.10(f) they were selected only after more than 1.5s.

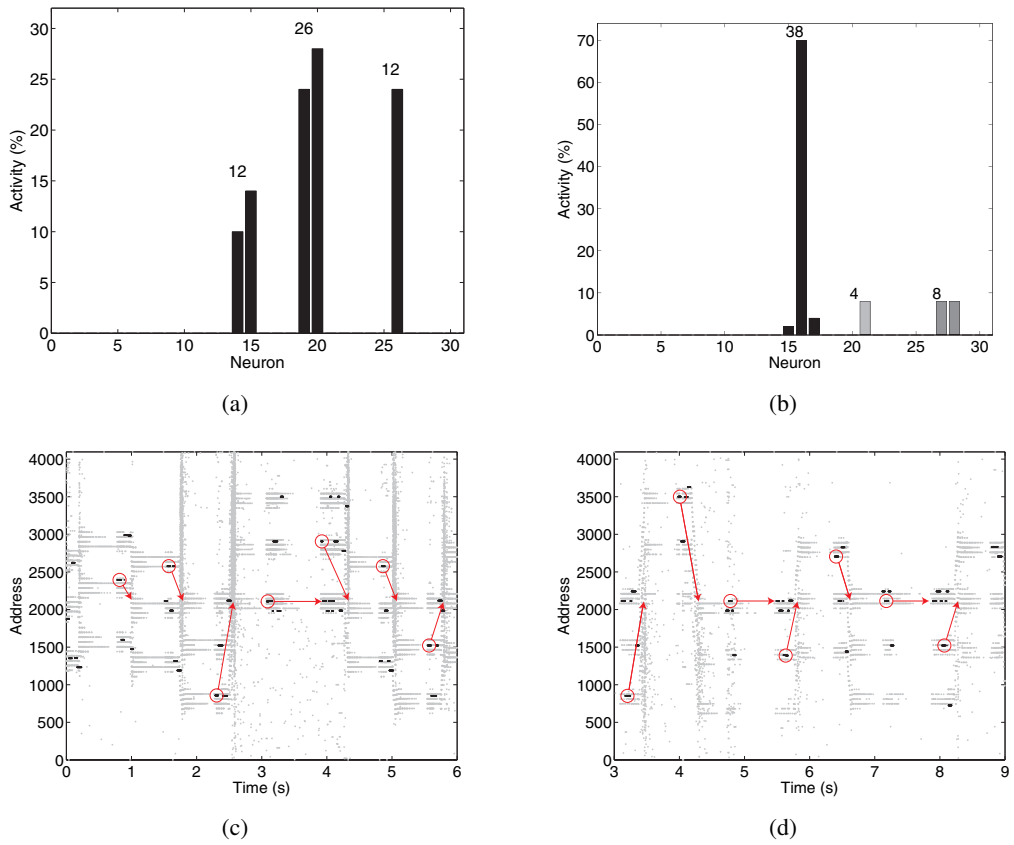


Figure 5.12: Overt attention scan path when the retina is stimulated with three circles. The stimuli for left and right columns are shown in Fig. 5.9(a) and Fig. 5.9(b) respectively. The top row shows the histogram of the pixel selection, as explained in Fig. 5.10. The bottom row shows the corresponding raster plots, in retino–centric coordinates. The fovea of the retina corresponds to the address 2080. During the saccade, the retina is moved such that after the saccade the selected pixel, highlighted by a red circle, is centered at the address corresponding to the fovea (red arrow). The selected pixel is not always the first pixel to burst, depending on the delays of the spike acquisition, and on the relative duration of the acquisition window ($\Delta t = 50$ms) and of the bursts. The experimental settings correspond to those of Fig. 5.10, with inhibitory time constant set to $V_{\tau inh} = 20$mV, to increase the burst duration.

the cases a single pixel with maximum activity can be selected from the activity monitored during the acquisition window.

The raster plots and the scan path generated with the "fast" and "slow" IOR settings are qualitatively similar because the random selection of the saccade target in the "fast" IOR configuration does not significantly change the statistics of the selection.

Fig. 5.12 shows the histograms and raster plots for overt attention with the "slow" IOR settings. As for the covert attention model experiments of Sec. 5.3.3, in the "baseline" condition the three circles are selected, and the difference in the number of time each pixel was selected can be attributed to the mismatch effects in both the retina and SAC pixels. Also in this case, in the "grey level" condition, the black circle is selected more often than the grey and light grey circles respectively. Fig. 5.12(c) and 5.12(d) show the raster plots of the SAC and retina activity in retinal coordinates; the three circles of the input image correspond to the three bands of activity. During the saccadic movements the position of the circles on the retina array changes, and the pixels that correspond to the trajectory of the circles on the retina array produce spikes, as can bee seen from the vertical stripes of activity. During the saccade the SAC does not receive any input from the retina, therefore it

does not produce any output. Immediately after the saccade landing, the retina jitters on its support, and the pixels corresponding to the new location of the circles on the retina array emit spikes, then the jitter stops and the retina does not produce any spike. After the saccadic suppression interval, the mapping to the SAC and the micro–saccades are re–enabled, and the target for the next saccade is selected.

The dynamics of the overt FOA scan path generated by the two–chip system depend on the interaction of the hardware and software components, and on the relative timing and delays of each operation.

The average delay from the generation of an event in the retina to the first spike generated by the SAC is about 10ms. This measure gives a rough estimate of the delay in the communication between the chips, and of the SAC processing time, including the synaptic integration, the WTA selection, and the $I\&F$ time to first spike.

Additional delays are introduced in the software component of the system. One source of delay is the time interval between the beginning of the micro–saccades and the beginning of the acquisition of address events from the SAC. The duration of the acquisition of the SAC address events is defined by the previously defined software acquisition window. After acquiring the address events and computing the target for the next saccade, and translating it into the appropriate commands for the PTU, the PTU motor commands are executed and the PTU moves after a delay determined by the communication with the serial port. Figures 5.12(c) and 5.12(d) show that the system performs roughly one saccade per second.

This overt attention experiment can be modified and optimized, for example by testing different software algorithms for the target selection of the saccade. Nevertheless the implementation described here allows the retina to foveate the different stimuli, with a frequency that depends on the input contrast.

The current implementation comprises a saccadic suppression mechanism of fixed duration of 500ms, independent on the real saccade duration. This guarantees that the input to the SAC is well separated from the events produced during the saccade.

The inherently asynchronous nature of the AER chips is not fully exploited in the current setup. The SAC does not control directly the PTU, instead a software algorithm translates the SAC address events into an appropriate motor command for the PTU. The software implementation involves the acquisition of the SAC address events for a fixed time frame. This strategy can introduce potential artifacts, specifically the selected target for the saccade depends on the relative timing between the spikes emitted by the SAC winner, and the beginning and length of the acquisition window. If more than one SAC pixel is active during the acquisition, an additional decisional stage has to be included to select an unique target. In the current implementation a “max” operation is used, but other strategies could be implemented, such as the selection of the first pixel that emitted a spike.

An asynchronous approach would be desirable for the read–out of the SAC activity, in particular the addresses of the winning neuron could be used directly to drive the PTU motors. A direct feed–back from the PTU to the PCI–AER for implementing saccadic suppression would also be desirable. In this hypothesized system the “slow” IOR settings with long bursts of activity corresponding to the winning pixels would be the best choice.

5.3.5 Covert and overt attention with moving stimuli

The final experiment was performed with freely moving stimuli in front of the retina. At the beginning of the experiment the retina is still, and a covert FOA scan path is produced. Subsequently the motor movements are enabled, and the PTU is used to foveate the selected stimulus. I first tested the system with two “smileys” drawn on a white paper, then with

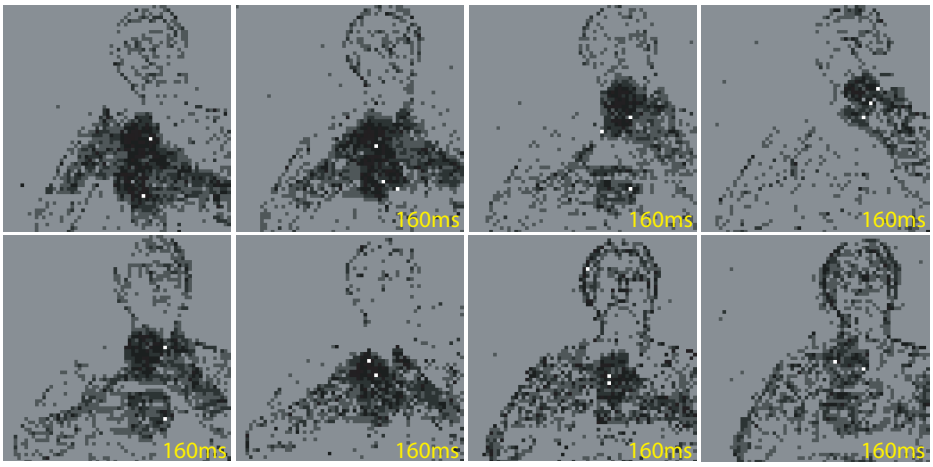


Figure 5.13: Covert attention with natural moving stimuli: screen shots of the recorded activity from the retina (black and grey pixels) and SAC (white pixels). The numbers on each screen shot correspond approximately to the time gap from the previous screen shot. The hands are strong attractors for the SAC selectivity, since they produce a strong circumscribed response in the retina. The arms edges are selected less often because the activity they elicit is more distributed. The edges of the head are selected rarely (second last screen shot).

hands in movements, and eventually with a person walking in front of the retina.

In case of the two smileys, as well as with moving hands, during the covert attention period, the SAC output oscillates between the locations of the two objects. Fig. 5.13 shows selected screen shots[1] for covert attention with moving hands. The SAC alternates between the two hands, while it selects only rarely the arms that are moving as well. This behavior confirms once more that stimuli with a dense and circumscribed region of activity, as the hands, have competitive advantage with respect to single edges, thanks to the lateral excitation (see Sec. 4.3.4). Fig. 5.14 shows the saccades of the retina tracking the two smileys: when the the paper is moved in one direction, the saccades center one of the two smileys in the middle of the retina, following the movement of the stimuli.

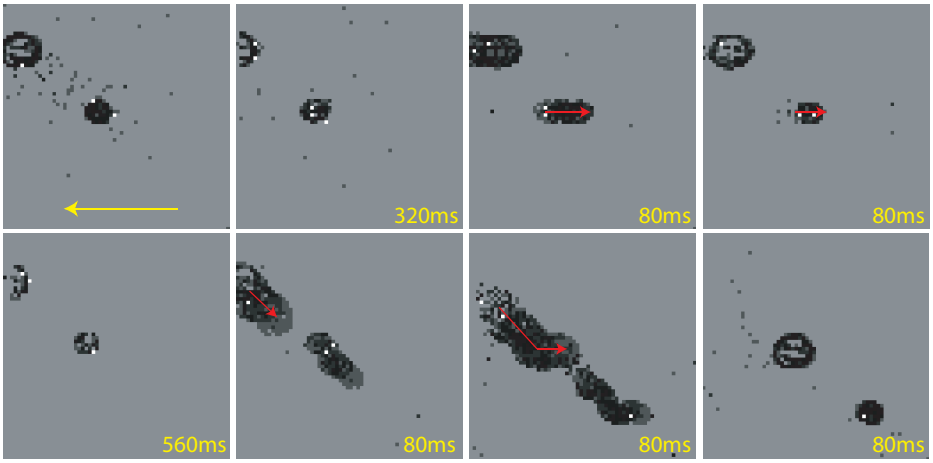
The speed of stimuli for which the system works is limited by the delays introduced by the software algorithm and by the communication between the software and the hardware components, as described in Sec. 5.3.4, and by the fixed duration of the saccadic suppression mechanism. After the saccade, the system waits until the end of saccadic suppression period to compute a new location for the next saccade: if the tracked stimulus moves too fast, it can disappear from the retina's field of view, before the system is ready to select the new saccade target.

Fig. 5.15 shows examples of sequences of target selection, saccade movement and landing, when the system is tested with a person walking in the field of view of the retina. The SAC selects pixels belonging to the person's contour and the system reliably follows the subject's trajectory.

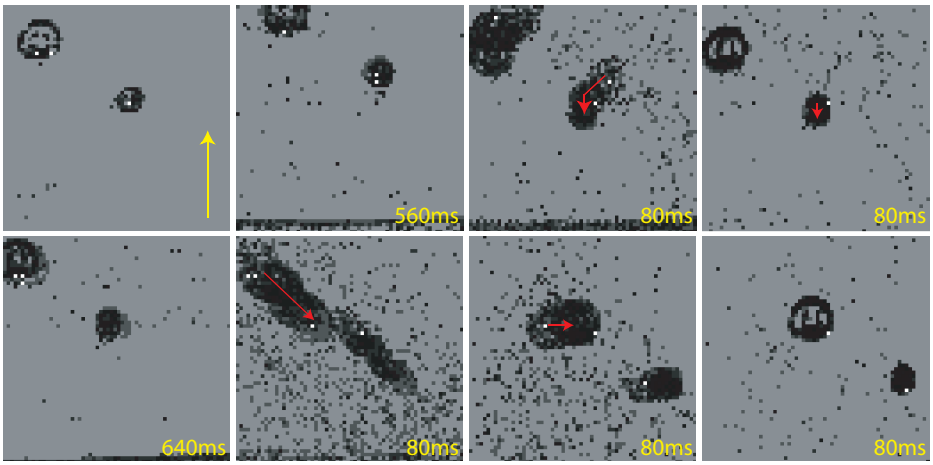
5.4 Conclusions

In this chapter I described the functionality of the SAC, as part of a basic selective attention system, comprising a neuromorphic sensor, the SAC as post–processing unit, and an actuator. I presented experiments that show how the system can select and track salient stimuli,

[1]The chip's activity is grouped into frames and displayed as a movie by a CaviarViewer software tool designed by T. Delbrück. In the experiments shown in this section, the frame duration has been arbitrarily chosen to be 80ms.



(a)



(b)

Figure 5.14: Overt attention with moving stimuli: the smileys are moved first to the left (a), then upward (b), as indicated by the yellow arrows. The screen shots show the sequence of movement, target selection, saccade (red arrow), and saccade landing. Both sequences show the foveation of the smaller smiley (upper row), followed by the foveation of the bigger one. The deflection of the saccade arrow in the second and third saccades is an artifact due to the Pan–Tilt Unit, that is not optimized for matching horizontal and vertical movement durations.

both covertly and overtly. When stimulated with real world stimuli, the system can select and follow the most salient moving objects.

5.4.1 IOR

The function of IOR in selective attention is a highly debated topic in the literature (Klein, 2000; Horowitz and Wolfe, 1998). The term was used first by Posner (1980), referring to the increasing reaction time for the visual processing of previously attended targets. From Posner's influential work, IOR has been interpreted as an inhibitory tagging of attended locations, that prevents the "reorienting" of attention towards such locations.

There is evidence for inhibitory tagging of multiple objects in visual search displays (Dodd and Pratt, 2006; Danziger et al., 1998), with IOR starting as soon as a location is chosen for attentional selection (Dodd and Pratt, 2006). The IOR build–up is masked by early attentional facilitation, it is revealed when attention moves away, and it decays with

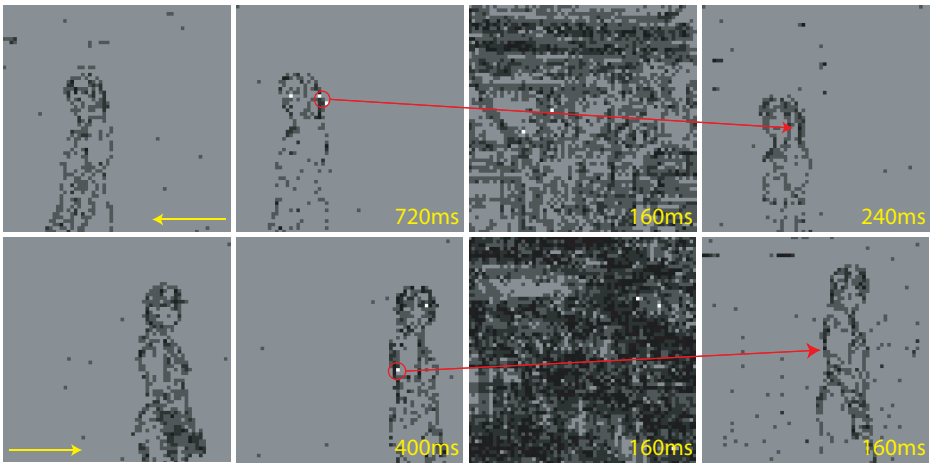


Figure 5.15: Overt attention with moving natural stimuli: two sequences of movement (yellow arrow), target selection (red circle), saccade movement and landing (red arrow), when a person is walking to the left (upper sequence) and to the right (lower sequence).

time. This is exactly the mechanism implemented on the SAC: as soon as one neuron wins, self–inhibition builds up, and when the neuron stops winning, it slowly decays, lasting for a relative long time.

This view has been challenged (see (Klein, 2000) for a review). Specifically Horowitz and Wolfe (1998) deny the existence of such short–term memory mechanism in visual search. Also the origins and neural mechanisms for IOR are still object of intense research and debate (Taylor, 2006).

To build a system for deploying attention, that can robustly operate in realistic scenarios, with a mixture of stationary and moving objects, a mechanism for deselecting the current attended position, or object, either voluntarily or not, is necessary.

Despite the straightforward implementation of IOR on the SAC, the experiments performed in this chapter are relevant in the debate about IOR, because they demonstrate the importance of the implementation of such a strategy in physical systems.

For covert attention in presence of static stimuli (see Sec. 5.2 and 5.3.3) IOR is necessary to disengage attention from the selected stimulus. The dynamics of the IOR, in particular the ratio between the duration of the activity of each winner and the duration of its suppression, strongly influences the attentional scan path. If the winners take a long time to self–inhibit and are depressed for a relatively short time, only few other pixels will have the occasion to fire before the strongest stimulus is able to compete again. Viceversa, if the build up of inhibition is fast relative to the suppression duration, more pixels will be comprised in the FOA scan path. In software implementations of the saliency map model, shape and size of the inhibitory region are sensible parameters, that influence the scan path. In the SAC the lateral diffusion circuits diffuse also the inhibitory current generated by the inhibitory synapse. The result is the spread of inhibition to the neighbors of the winner. In the SAC current implementation lateral diffusion of the inhibitory current decays exponentially with distance, and the space constant decreases with increasing amplitude of the inhibitory current. The experiments of this chapter and of Sec. 4.3.4 show that there is a limited range for tuning the lateral diffusion. To be able to set an arbitrary size of the inhibition diffusion, more elaborate diffusion networks should be implemented, at the cost of higher complexity and size of the circuits. An alternative strategy for implementing lateral connectivity is to exploit the linearity of the DPI synapse circuit, and use the local inhibitory DPI synapse of each pixel to integrate the activity of the corresponding I&F neuron, and of its first and

second order neighbors. This solution improves the flexibility of the IOR mechanism, and requires only two additional transistors per pixel and one extra voltage bias.

The experiment described in Sec. 5.3.2 shows that the IOR is needed also when the stimuli move, to counteract hysteresis. The hysteretic mechanism associated with lateral excitation were introduced to favor the tracking of a selected stimulus over the selection of distracters. When a stimulus is selected, a fraction of the hysteretic current diffuses to the neighbors. If the stimulus moves smoothly, it stimulates the neighbors of the current winner, that receive in addition the diffused hysteretic current. The diffused current is exponentially smaller than the hysteretic current of the winner. Also in such a situation the system would benefit from a more elaborate implementation of the diffusion circuits.

IOR is necessary as well for overt attention. In experiments with static stimuli as those described in Sec. 5.3.4, it is necessary to disengage attention when the most salient stimulus has been foveated. It is still matter of debate if IOR is a mechanism that inhibits the re–orienting of attention to previously selected location, in order to facilitate visual search and inspection of scenes (Klein, 2000; Horowitz and Wolfe, 1998; Dodd and Pratt, 2006). In such a case a retinotopic saliency map would be inadequate, and a world–centered map would be necessary (Posner and Cohen, 1984; Klein, 2000). Morgan et al. (2005) assert that IOR is related not only to locations, but also to objects. In such a case an inhibitory tag is attached to each attended stimulus, and IOR has to follow the movement of the inhibited stimuli. The physical implementation of the attentional system poses practical question such as the need for coordinate re–mapping and system reference frames (Colby and Goldberg, 1999). With the hardware system I developed one could exploit the PCI–AER mapping functionality to map the retina activity in world–centered coordinates, or in object–centered coordinates, and investigate different hypotheses with real world stimuli.

5.4.2 Attentional tracking

In presence of natural scenes perceptive systems perform a combination of shifts of attentional selection and attentional tracking, with the scope of capturing information about the visual environment, and inspecting the selected targets.

Previously proposed hardware systems modeling attentional selection implement also forms of attentional tracking (Brajovic and Kanade, 1998; Fish et al., 2004; Horiuchi and Koch, 1999). Two systems that integrate attentional selection and tracking in analog VLSI have been proposed by Brajovic and Kanade (1998) and by Fish et al. (2004). Such chips have two modes of operation: in the "selection" mode WTA competition is enabled in the full pixel array of the chip, and a stimulus is selected as target. In the "tracking" mode the WTA competition is enabled only in a small tracking window centered around the position of the selected target, and all other stimuli measured by pixels outside such window are ignored. The WTA competition selects the next winner among the stimuli in the tracking window, and the window is centered on the position of the newly selected target. For implementing robust tracking, the system relies on the assumption that typically only the initial stimulus selected as a target during the "selection" mode is present in the tracking window. This approach prevents any stimulus outside the tracking window from shifting attention from the chosen target.

The system selects a new target only if the tracked stimulus exits the field of view of the chip, or if an external observer switches the operating mode of the chip. Although useful in tracking applications, this approach is not biologically plausible, since there is evidence that attention enhances the processing at the attended location (Lee et al., 1999), and is not a filter that blocks everything outside the region around the attended stimulus (see (Ambinder

and Simons, 2005) for a critical review on attentional capture and inattentional blindness). Additionally such VLSI system lacks an automatic deselection mechanism such as IOR, that allows shifts of attention for the scanning of visual inputs.

Horiuchi and Koch (1999) and Indiveri (1999) proposed a more biologically realistic approach. Instead of filtering out everything except for the target and its immediate surrounding, the saliency of the target and its surrounding are enhanced with hysteresis and lateral excitation. These systems were designed explicitly for implementing attentional tracking, so they do not perform attentional shifts unless a more salient stimulus appears in the field of view, or when the tracked stimulus disappears.

Similar to the approach described in (Horiuchi and Koch, 1999; Indiveri, 1999), the chip proposed in this thesis comprises both attentional tracking and attentional shifts with the inhibition of return mechanism. In these systems the two mechanisms interact dynamically; depending on the chip's bias parameters there is a different balance of attentional shifts and tracking. With weak inhibition, strong hysteresis, and lateral excitation, the chip is optimally tuned for "tracking" mode. Conversely with strong inhibition and weaker hysteresis, the chip produces repetitive FOA shifts and is optimally tuned for "searching" mode.

In the implementation proposed by Horiuchi and Niebur (1999) the chip can be utilized either in "tracking" or "searching" mode; an external module needs to be used to set the parameters for the desired operating mode. The control of IOR, hysteretic current and lateral excitation, can be interpreted as a voluntary top–down biasing in favor of tracking when the predefined task–dependent stimulus is selected in "search" mode. Also in this implementation the "tracking" mode ends only when the target disappears, or a more salient target appears.

Also in the SAC implementation it is possible to use an external module, to perform the transition between the "search" and the "tracking" mode. However, the philosophy of this project was to implement a compact and autonomous device capable of scanning the visual input autonomously. Specifically I implemented a system that sequentially selects all of the salient input stimuli. To model the combination of attentional shifts and tracking of perceptive systems, the device is able to disengage automatically from a tracked target. In principle, depending on the parameters configuration of the SAC, the selected stimuli are deselected with different inertia, favoring a mixture of scanning and attentional tracking. The experiments performed in Sec. 5.3.2 show SAC attentional tracking of a moving stimulus: with the resistive network described in Sec. 3.3, the space constant for lateral diffusion of the hysteretic current is quite small, and for large values of hysteresis, that should favor tracking, the chip locks the system on the selected pixel. In such a case IOR is necessary to suppress the winning pixel, but tracking works only in absence of distracters. The tracking behavior of the system developed would benefit from the implementation of more elaborate circuits for lateral diffusion, or on the implementation of a smooth pursuit mechanism (Horiuchi and Koch, 1999).

When human subjects freely inspect a visual scene with a combination of stationary and moving stimuli, the eye movements show a mixture of saccades and smooth pursuit movements, that follow moving targets. The relationship between smooth pursuit movements and attention, and the behavioral relationship with respect to saccades are still under debate (Horowitz et al., 2004; Khurana and Kowler, 1987). Nevertheless there is evidence that smooth pursuit is linked to saccades (Gardner and Lisberger, 2002), in that smooth pursuit follows the object foveated by saccadic movements, and that attention is involved in the selection of the smooth pursuit target (Ferrera and Lisberger, 1995).

5.4.3 Relevance of the two–chip system implementation

The two–chip visual system described in this chapter is the first realization of a full–custom AER–based perceptive system, capable of sequentially selecting relevant regions of the input stimuli. The SAC performs a non–trivial processing on the output activity of a neuromorphic silicon sensor. The system design incorporates biologically inspired computing principles. The processing of sensory data is asynchronous and event–driven, implementing both efficient computation and communication (see Sec. 1.1). The specific computation performed by the SAC extracts relevant information via a cooperative/competitive mechanism that takes into account the relative context of the input, rather than the absolute value of each component (pixel). The implementation of an adaptive mechanisms at the synaptic level, reproducing the short–term depression phenomenon observed in real synapses, results in a high–level property of the system, that makes the system sensitive to transient stimuli, while suppressing stationary stimuli.

The experiments described in this chapter demonstrate that the multi–chip system developed can perform reliable visual inspection of the environment. It is therefore a potentially efficient and robust mean for the implementation of autonomous artificial perceptive systems, for example supporting robotic navigation.

The realization of a system with an actuator that closes the sensory–motor loop by orienting the sensor towards salient stimuli shows that the system can respond with an appropriate action to the sensory stimuli. The experiments of Fig. 5.10 show that the SAC output defines an univocal winner whose activity is stable for tens of milliseconds, that can be directly used to control the actuators.

The analysis of the behavior of the system when confronted with real world stimuli shows the effect of the dynamical interaction and the relative roles of its different components, separately analyzed in Chap. 4. Specifically the experiments of Sec. 5.2 highlighted the relatively weak contribution of the lateral diffusing circuits to the overall behavior, demonstrating that the system would benefit from the implementation of a stronger form of cooperation.

The implementation of the two–chip system has proven that the IOR mechanism is necessary for the functioning of the system, and that the implementation of the smooth pursuit mechanism could improve the tracking behavior of the system. The phenomenological emulation that includes in an automatic bottom–up fashion mechanisms that can have diverse origins is sufficient for the implementation of a compact and flexible device reproducing some of the basic properties of attentional selection.

Chapter 6

Discussion

6.1 Relevance of the work described in this thesis

Research on neuromorphic circuit design for the emulation of biomimetic computational functions is relevant for understanding the computational strategies used by the brain to overcome constraints such as limited space, wiring, power and precision, imposed by the physical realization of computation (Douglas et al., 1995). In neural systems these constraints led to the evolution of robust computation based on conceptually different principles from those of classical digital computation, far more efficient in solving ill–posed problems and extracting reliable information from noisy and ambiguous data (Mead, 1990; Douglas et al., 1994).

The goal of neuromorphic engineering is the development of new technology for implementing computational principles based on the same principles as the neural substrate. Neuromorphic engineering is an heterogeneous field,with research areas covering a wide range of aspects, from modeling specific properties of neurons and synapses, to studying the computational properties of networks of interconnected neurons, to the development of compact VLSI sensory devices for commercial and biomedical applications (Lichtsteiner et al., 2006a; Sarpeshkar, 2006; Boahen, 2005).

This thesis touches on the different aspects of the neuromorphic approach. It begins with the description of a new circuit for modeling synaptic currents and emulating specific properties of synaptic transmission. It then describes a WTA circuit that reproduces competitive and cooperative aspects of computation emerging from the recurrent interaction between neurons (Douglas et al., 1995, 1999).

These elements are combined in the realization of a device that implements selective attention: a complex high–level function.

The work described in this thesis merges basic research, such as the design of models of synaptic transmission, with applied research, through the effort of implementing a custom multi–chip system operating in real time with a well defined function, which can be applied to practical problems, for example robotic navigation or visual scene inspection.

The path required to implement, test and build a neuromorphic multi–chip system comprises the study and design of the hybrid analog/digital neuromorphic circuits, as well as the development of communication and control infrastructure to interact with the chip and connect the various modules together. An important aspect of my work was to integrate all components of the infrastructure, merging at various levels of abstraction the software and the hardware needed for the system setup, which were developed throughout the years within different frameworks.

6.1.1 The silicon synapse

The new silicon synapse circuit developed in this work is extremely relevant for neuromorphic research in general. The circuit reproduces a good approximation of the currents generated by presynaptic action potentials on the postsynaptic neuron's membrane. This faithful hardware model of current dynamics increases the similarities between silicon and biological synaptic transmission. The exponential time course of the postsynaptic currents implements synaptic summation, an essential property of synaptic transmission, observed in real neurons and included in computational models (Destexhe et al., 1998). The additional circuits proposed in Chap. 2 extend the basic synaptic functionality by including the phenomenological implementation of additional synaptic properties, such as NMDA voltage–gating and conductance–based behaviors. Such circuits enrich the ensemble of computational primitives that can be emulated on silicon, in a unified framework that can comprise all of them in a single compact circuit. When included in large recurrent neural networks, it is possible to study the computational effect of such primitives in real time. An important primitive introduced with the DPI design is synaptic homeostatic plasticity (Turrigiano, 1999). It is one of those adaptive mechanisms, observed in real neural populations, that can be used to reduce the intrinsic inhomogeneities of the computational substrate (Mead, 1990) and is now implementable on silicon. Implementation of homeostasis in neuromorphic chips will improve the stability, robustness, and mismatch tolerance of networks in the face of tuning of circuits parameters. It will render chips automatically adaptive to chronic changes in their environment, from variations due to temperature or power supply fluctuations, to the change of input range caused by failure or inclusion of new input devices.

6.1.2 The selective attention chip

The majority of devices designed in the framework of neuromorphic engineering until recently focused on the sensory interfaces of brains. My interest was in the emulation of neural systems that process such signals; specifically I focused on selective attention, a well known strategy exploited in visual perception for overcoming the problem of limited parallel processing capabilities (Itti et al., 2001). Selective attention guides the selection of important regions of the input stimuli for sequential allocation of computational resources. It is an essential function for perceptive systems, and its implementation in artificial systems is crucial for perceptual tasks.

The experiments described in this thesis show that the SAC can be reliably used for the implementation of multi–chip selective attention systems. Specifically, the hardware device I developed has a dual relevance: it can be used as a tool for basic research, and as component of artificial systems used in practical applications.

The relevance of the hardware implementation is in providing a tool for exploring different hypotheses about the mechanisms involved in stimulus dependent attentional selection, including some simple forms of top-down influence, in real time, and using real world stimuli.

The SAC can perform computationally intensive operations, with complex dynamics, in real–time. It has been designed in such a way as to allow the use of multiple instances of the same chip in hierarchical multi–chip systems. The various instances of the SAC could implement the functionality of the software modules that implement spatial and feature competition in the Itti et al. (2001) model.

The flexibility of the system derives from the use of the AER communication protocol. The AER infrastructure I assembled (see Appendix C), allows the full automation of some

experimental procedures, making extensive parametric analysis possible (e.g. see Chap. 4).

The constraints and limitations imposed by the physical realization of the selective attention system can guide the design of feasible and plausible strategies for the functional implementation of the perceptual attentional selection observed by experimentalists. In this context, the experiments with the SAC embedded in a minimal selective attention model with real dynamic stimuli (Chap. 5) demonstrated the necessity of a mechanism such as IOR in a purely bottom–up system. I showed how this form of deselection of the attended object can be useful for visual search. Other experiments confirmed the essential role of lateral excitatory connectivity, or cooperative computation, in decreasing the effects of inhomogeneities in the computational substrate, improving the system's overall performance. Yet another form of recurrent connectivity, self–excitation, was shown to improve the discriminability of stimuli. These properties are responsible for the system's "high–level" functionality. Examples of high–level operators include short–term memory, provided by the self–excitation, and the competitive advantage given to regions, rather than to isolated pixels, provided by lateral excitation. This is a practical example of how the mechanisms adopted for coping with limitations in the physical substrate at the same time lead to the extraction of meaningful information from the input signals, which is the essence of neural computation (Douglas et al., 1995).

Some of the characteristic behaviors observed in psychophysics experiments such as IOR, or a bias towards short saccades, have been attributed either to top–down effects or to the mechanical properties of the eye movement "plant"; such topics are still a matter of debate and research. One of the scopes of the work described in this thesis was the realization of a device that can be used in practical applications; a phenomenological implementation that includes in an automatic bottom–up fashion some of the mentioned mechanisms is necessary and sufficient. A striking example is the recent quantification of the contribution of bottom–up saliency for guiding unconstrained overt attention. The study reported by Carmi and Itti (2006) shows causality between bottom–up saliency and saccade selection, especially when motion contrast is included in the computation of the saliency map. The inclusion of short–term depression in the input synapses of the SAC makes it intrinsically selective to variations of the input, therefore to motion and to transient stimuli. Regarding the contribution of other visual features, for example contrast or orientation, to the computation of the saliency map, especially for still images, the debate is still open wether there is causation or simple correlation, and to what degree, between such features and attentional selection (see (Einhäuser et al., 2006) for a review). For the purpose of building autonomous devices like the SAC, it is meaningful to causally include in the computation of the saliency map features that are correlated with attentional selection, because it is probable that such features are linked to the selection of behaviorally relevant objects. It is also reasonable to suppose that during evolution, correlations between stimulus features and stimulus behavioral relevance became more and more implicit in the system, becoming part of stimulus–driven visual selection; e.g. color might be relevant when picking a fruit among green leaves (Carmi and Itti, 2006).

Another point worthy of discussion which emerged from studies of the bottom–up contribution to attention is the finding that the selection of stimuli with high computed salience is above chance, but the probability decreases for the selection of the maximum of the saliency map (Itti, 2005a). This observation questions the role of the WTA competition as a computational mechanism for the selection of attentional targets from the saliency map. The scenario proposed by (Itti, 2005a) is that the saliency map is a sparse representation of the input, indicating the most salient stimuli, then a top–down or random mechanism selects an attentional target from among them. However, the SAC implementation of atten-

tional selection via WTA competition is still useful for stimulus selection in the absence of top–down influences on an autonomous device; as observed above, the peaks of the saliency map are related to relevant stimuli, and sequentially selecting them in order of decreasing saliency is an effective strategy for the exploration of the environment.

The design of chips such as the SAC leads to the realization of compact, low–power, portable devices that can be used in practical applications. Silicon sensors such as retinae and cochleae are nowadays well characterized, and begin to be robust enough for use outside of the laboratory. Regarding potential commercial applications, a promising direction for these devices is their use in prosthetic systems.

The SAC goes a step beyond this approach: it implements a crucial post–processing phase for sensory signals, using the same technology. Its AER based design makes it suitable for receiving signals from multiple AER silicon sensory devices. In Chap. 5 I demonstrated that when connected to a silicon retina the SAC can orient the retina towards salient and moving stimuli. With the evolution of a compact stand–alone AER infrastructure, this selective attention system could be easily mounted on a robot.

6.2 Outlook

I believe it is important to pursue research in the field of silicon emulation of neural computation by modeling properties of the basic components of the neural substrate, and by studying the effect of such properties on computation at the network and system level. It is also important to begin applying the resulting technology to the development of compact devices for practical applications in the real world. Applications where low power consumption and real time interaction with the real world are essential requirements can benefit from the use of neuromorphic devices; potential applications are in autonomous robotics and navigation, but also brain–machine interfaces, prostheses and implantable devices, and medical image processing. From this perspective, two main projects could carry on the work described in this thesis. An important first step would be the study and implementation of circuits of the control loop for synaptic homeostasis. This adaptive mechanism is crucial for making VLSI devices more robust to the intrinsic mismatch of their computational units. In the context of implementation of selective attention, apart from further improving the engineering aspects of the hardware implementation, a possible development would be the design of different sensors for input to the SAC. Recent psychophysics experiments suggest that motion is a strong predictor for the deployment of attention; the re–design of the optic–flow sensor proposed by Stocker and Douglas (1998) in context of AER would make it suitable for generating a motion–based saliency map for the SAC.

In conclusion, the work presented in this thesis is relevant for the advance of neuromorphic research at different levels. The proposed synapse circuit proposed is a step forward in silicon emulation of computational primitives typical of the neural substrate. For the purpose of building a device that can be included in artificial perceptive systems, the SAC implementation with all its limitations is nevertheless useful and significant. It represents a first step in the evolution of devices that can be used for practical applications.

Appendix A

Linear–Threshold Units Winner–Take–All simulations

This appendix lists the results of software simulations of a simple recurrent network showing Winner–Take–All computation.

A.1 Recurrent WTA Networks

Many studies address the issue on how neurons are connected in the brain. The main part of these researches focuses on visual cortex, trying to figure out what kind of circuits are involved in visual perception (Anderson et al., 1998).

The purely feed–forward structure, with high parallelism, proposed by Hubel and Wiesel (1977), accounts for the structure of simple and complex cells receptive fields, thus it accounts for cells selectivity to different features of visual stimuli. However, studies on neuroanatomy pointed out that most of cortical connections are made locally and that neocortical neurons avoid long connections (Douglas et al., 1996; Martin, 2002). Furthermore in the feed–forward model the degree of cells orientation tuning is highly dependent on the contrast of the input and is very susceptible to noise (Douglas et al., 1999).

An additional computation is required to explain the highly reliable response of cortical neurons.

The hypothesis exploited by our research is that the parallel, feed–forward, flux of information is analyzed by mean of recurrent circuits that amplify the most effective input and suppress the weakest (WTA networks). This helps to sharpen the selectivity also for low contrast stimuli and high levels of background noise.

Many different configurations for recurrent connectivity have been proposed (Ben-Yishai et al., 1995; Salinas and Abbott, 1996; Hahnloser et al., 2000).

Ben-Yishai et al. (1995) developed a model in which orientation tuning is generated primarily by recurrent rather than feed–forward connections. The model connects Linear–Threshold Unit (LTU) neurons and uses a cosine function for synaptic weights; all cells send and receive inputs from a global inhibitory neuron. Recurrent amplification enhances orientation selectivity of the cells. The contrast level of the input modulates the amplitude of the response, but doesn't affect the tuning.

Salinas and Abbott (1996) use LTUs to model direction selectivity in posterior parietal neurons. In this case the synaptic weight function is a DOG (difference of gaussian), that includes both excitatory and inhibitory interactions. The cell direction tuning is not affected by an added constant input, that in this model encodes for head direction: it will only modulate the output amplitude.

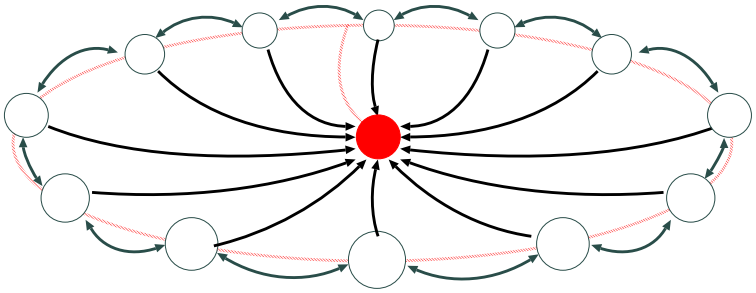


Figure A.1: Ring of neurons: excitatory connections are represented in black, inhibitory connections are in red.

Hahnloser et al. (2000) realized a cortex–inspired silicon circuit consisting of fifteen excitatory LTUs connected to a global inhibitor and linked by nearest–neighbor and second–nearest neighbor connections. Such network shows digital selection of the strongest stimulus (WTA competition) and graded analog response. The set of active neurons is determined by the connectivity pattern and by the input; the amplitude of the response depends on the background activity level, implementing gain modulation. A property of such network is multi–stability: when two isolated inputs are similar either one can be selected, but once the selection has been made it remains stable, even in face of small fluctuations of the inputs.

In this Appendix I show the results for simulating a 'simple' WTA, obtained reducing lateral connections to the nearest–neighbors and using one global inhibitory cell. The simulations are performed with Matlab, using LTU as model neurons.

A.2 WTA performances

The network is composed by a ring of excitatory neurons that project their output to a global inhibitory neuron, that in turn will inhibit the ring (Fig. A.1); w_{ei} is the strength of excitatory connections versus the inhibitory neuron, w_{ie} is the strength of inhibitory connections versus the excitatory neurons and w_e is the strength of input synapses. The excitatory neurons have first–neighbor lateral connections, w_{el}.

Each excitatory neuron receives an input current x_j from outside, the corresponding output frequency is $y_{e,j}$ (eq. (A.1)a). The inhibitory unit receives inputs only from the ring, its output is y_i (eq. (A.1)b). We model the dynamic of the neural response with a first order approximation (Dayan and Abbott, 2001), using different time constants for excitatory and inhibitory units.

$$\begin{cases} \frac{dy_{e,j}}{dt} = -\frac{y_{e,j}}{\tau_e} + \frac{1}{\tau_e} f(x_j w_{e,j} - w_{ie,j}) \\ \frac{dy_i}{dt} = -\frac{y_i}{\tau_i} + \frac{1}{\tau_i} f(\sum_j w_{ei,j} y_{e,j}) \end{cases} \tag{A.1}$$

LTU output is equal to the input if it is positive, or above a given threshold, zero otherwise (eq. (A.2))

$$f(x) = max(0, x) \tag{A.2}$$

The network was characterized for varying inputs and strength of lateral excitatory and inhibitory connections.

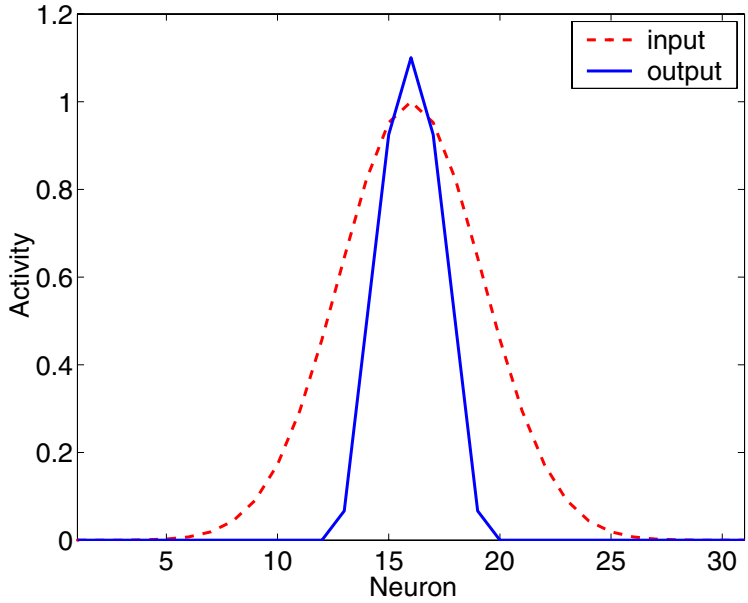


Figure A.2: Sharpening: plot obtained for $w_{ie} = 0.45$, $w_{ei} = 0.45$ and $w_{el} = 0.5$. The active neurons are less than the stimulated, because the ones receiving the weakest input are suppressed.

A.2.1 Suppression of less effective stimuli

The main feature of WTA network is to sharpen the selectivity of cells, suppressing their weakest inputs (Fig. A.2). This property accounts also for the ability on detecting a "target" feature among multiple "distractors", or noise.

Depending on the weights, this network can act as a strong WTA, that suppresses the response of all the neurons except the one receiving the strongest stimulus, or as a soft WTA, that allows more than one neuron to be active at the same time.

Suppression was measured as the difference between the maximum output frequency, ν_{max}, and the mean activity of all the "non–winner" neurons, $\nu_{mean,non-winner}$, normalized respect to the mean activity of the network, ν_{mean}:

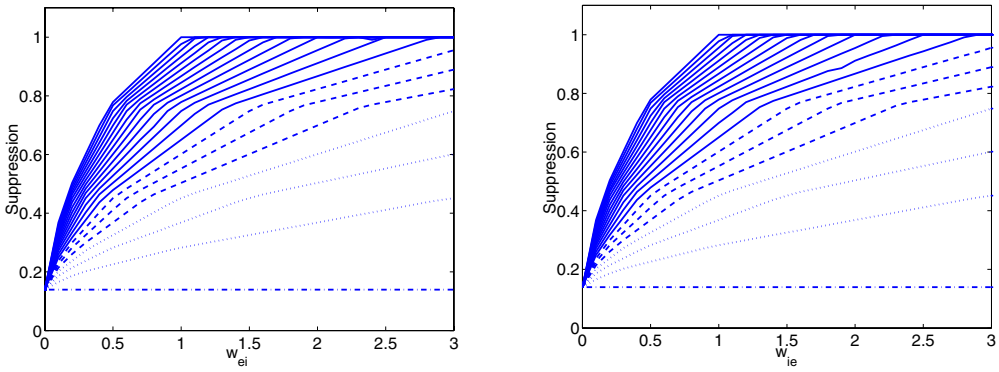
$$Suppression \doteq \frac{\nu_{max} - \nu_{mean,non-winner}}{\nu_{mean}} \tag{A.3}$$

The first results (Fig. A.3) refer to the case without lateral coupling: parametric curves respect to w_{ei}, measured as functions of w_{ie}, are equal to parametric curves respect to w_{ie}, measured as function of w_{ei}. This points out that the the effect of the two synaptic strength is symmetrical.

These plots show three different regions: if $w_{ei} > 2$ and $w_{ie} > 2$ the curves reach the maximal suppression, if $w_{ei} < 0.6$ and $w_{ie} < 0.6$ they do not, and for intermediate values ($w_{ei} > 0.8$ and $w_{ie} > 0.8$) they change slope.

For this reason three regions of network behavior can be distinguished: High coupling (high inhibition), low coupling (small inhibition) and intermediate region.

I measured the effect of lateral coupling in these three distinct regions, fixing the value of inhibition (Fig. A.3). Lateral coupling enhances the response of active neurons and of their neighbors, increasing the activity of excitatory units, this leads to a stronger activation of the inhibitory circuitry. These two effects compete, in the high coupling regime suppression decrease for increasing w_{el}, this means that excitation becomes stronger than inhibition. On the contrary, in the low coupling regime suppression increase for increasing



(a) (b)

Figure A.3: Suppression, parametric curves, in absence of lateral excitation: the measure is obtained computing the difference between the maximum output frequency and the mean activity of all the "non–winner" neurons, normalized respect to the mean activity of the network. As (a) and (b) are identical, the suppression is equally dependent on w_{ei} and on w_{ie}.

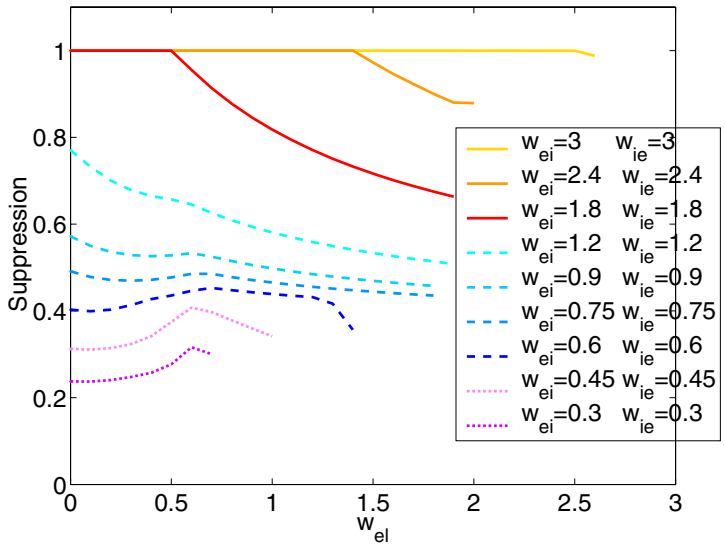


Figure A.4: Suppression, parametric curves, for varying inhibition: few curves are drawn for high coupling, low coupling and intermediate values of inhibition. As inhibition weakens, the value of lateral excitation that cause instability of the response decreases, as underlined by the shorter curves.

w_{el}, therefore inhibition becomes stronger than excitation.

A.2.2 Hysteretic behavior

Multi–stability property of the simulated WTA can be uncovered by hysteretic behavior. When two inputs are similar either of them can be selected, but once one of them is chosen the network remains stable even in face of small fluctuation of the input. In other terms, if a stimulus is winning the competition for saliency it has a competitive advantage over the other stimuli. When a new stimulus with intensity equal to the winning one is presented, it will not win, unless it becomes stronger than the first stimulus by a set amount (Fig. A.5).

As described in the main text (see Chap. 3 and 4), hysteresis is an important feature of the selective attention chip.

From the circuit point of view, it stabilizes the network and increases its resolution; for

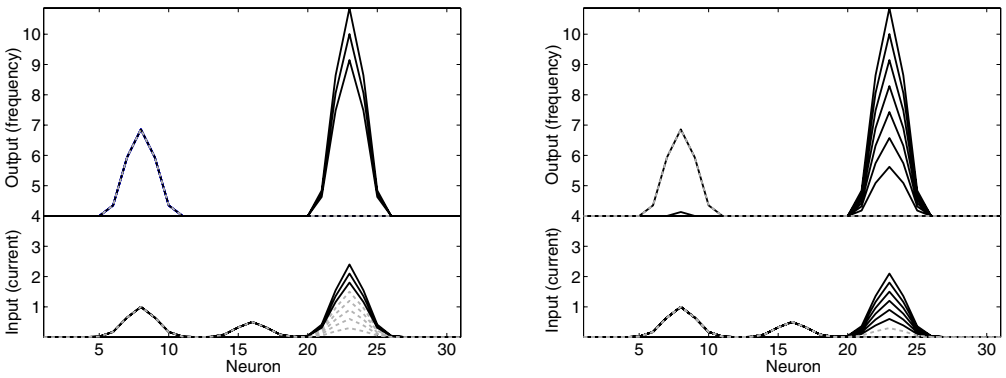


Figure A.5: Hysteresis: plot obtained for $w_{ie} = 0.45$, $w_{ei} = 0.45$ and $w_{el} = 0.87$. (a) The third gaussian input has to become higher than the first one to change the center of mass of the network; (b) to switch again to the first winner the third input has to become lower than the first one.

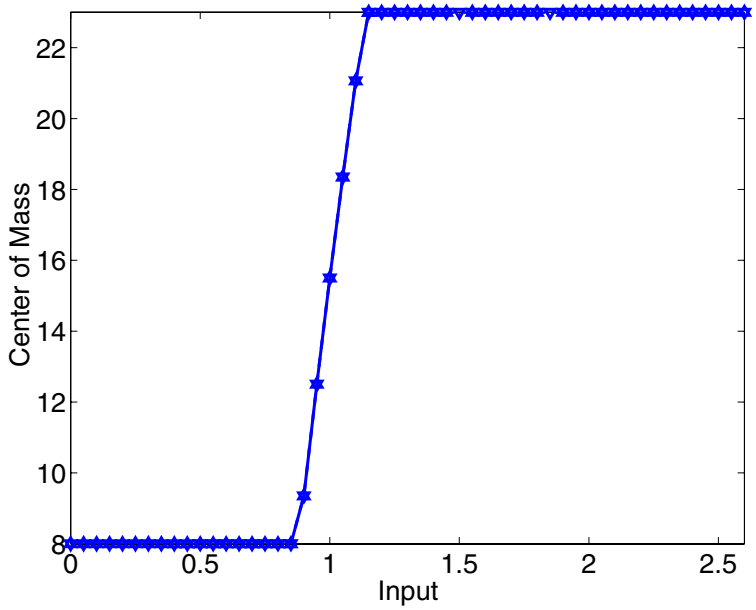


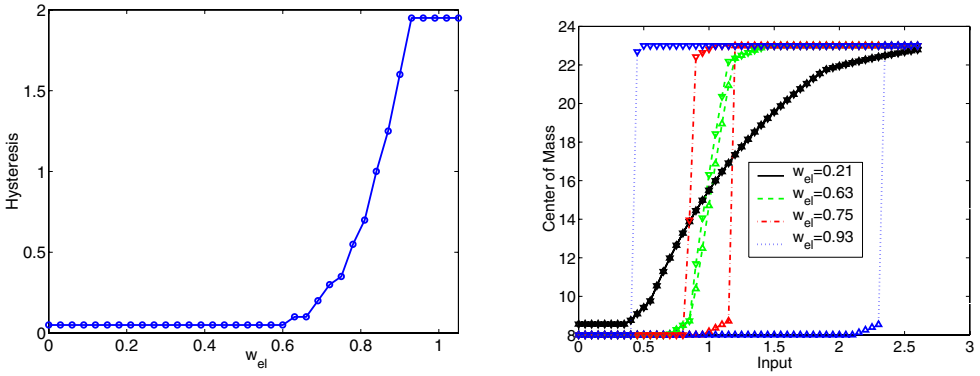
Figure A.6: Hysteresis, high coupling: Center of mass for a fixed lateral excitation $w_{el} = 0.99$ and high inhibition, $w_{ei} = 2.7$, $w_{ie} = 2.7$. Even varying the lateral coupling strength no hysteretic behavior is shown.

the implementation of attentional mechanism, it can account for 'tracking' a pattern even while its attributes are changing, because when a salient pattern changes in one or more of its attributes, some competitive advantage is passed on to the changed pattern.

I measured hysteresis in two regions (high and low coupling) and for intermediate values of inhibition.

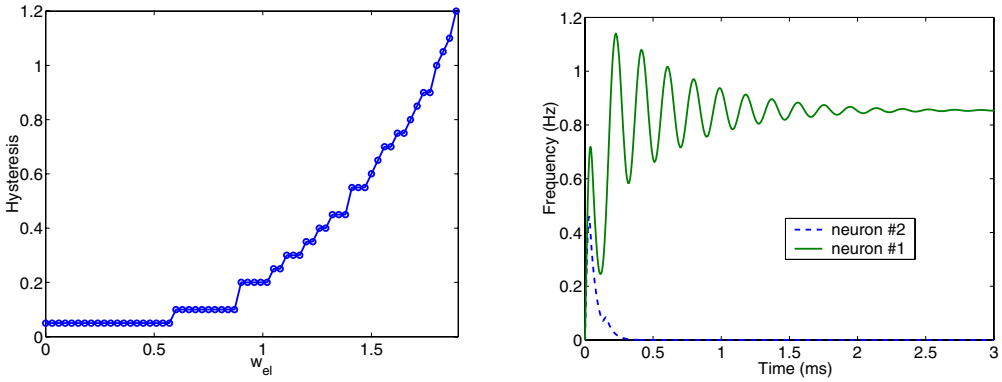
To measure hysteresis I calculated how the center of mass of the network's activity changes with the input, for different values of the parameters. The input is composed of three equidistant gaussian bumps, each centered on a different neuron, one of them changes its amplitude from below to above the normalized one, and back.

In the high coupling regime (i.e. high inhibition, $w_{ei} > 2$, $w_{ie} > 2$) the network does not show hysteresis (Fig. A.6). In the low coupling regime the hysteretic behavior depends on the lateral coupling w_{el}. If lateral coupling is small (i.e. $w_{el} < 0.8$) no hysteresis is shown,



(a) (b)

Figure A.7: Hysteresis, low coupling ($w_{ei} = 0.45$, $w_{ie} = 0.45$): (a) measure of the amplitude of hysteretic cycle, for low values of lateral excitation no hysteresis is shown, then it increases and, for really high w_{el}, the flat curve represents an "infinite" hysteresis, where the network doesn't switch to a second winner; (b) center of mass for some fixed lateral excitations.



(a) (b)

Figure A.8: Hysteresis, intermediate values of inhibition ($w_{ei} = 1.05$, $w_{ie} = 1.05$): (a) measure of the amplitude of hysteretic cycle, varying lateral excitation; (b) temporal response of the winning and the second strongest neuron, for $w_{el} = 1.8$; increasing lateral excitation the response oscillates before reaching the steady state.

as lateral coupling increases hysteresis appears and spreads with increasing w_{el}. As shown in figure (Fig. A.7), for low lateral excitations the response of neurons is independent on the previous state of the net, then there is a region where the winning neuron has an advantage on the others and with even stronger w_{el} the net will not switch to an other winner. For intermediate values of inhibition the hysteresis depends again on the lateral excitation, but with a different slope (Fig. A.8(a)). In this case we noticed that convergence time for the steady state response increases, because for the first few milliseconds the output frequency oscillates (Fig. A.8(b)). With the same value of lateral excitation, hysteresis increases as inhibition decreases.

The results presented refer to a network composed of 31 excitatory neurons; I have also measured how the hysteretic cycle varies depending on the number of neurons in the ring.

As long as the ring widens the hysteresis spreads (Fig. A.9): adding neurons has the same effect as increasing the strength of synaptic weights w_{ei} and w_{ie}, this is because the input is scaled with the number of neurons and more cells are stimulated; the effect is that the inhibitory neuron receives a stronger input and in turns it sends back a stronger

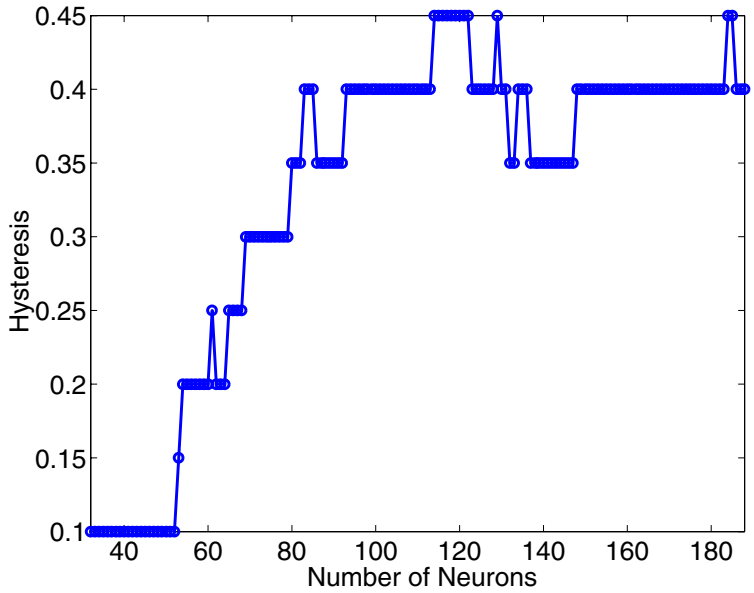


Figure A.9: Hysteresis, for increasing the number of excitatory neurons in the ring: the amplitude of hysteresis increases, but the effect is weak and affected by approximation errors.

inhibition to the ring. Anyway the result on hysteresis is weaker than, for example, the one obtained by increasing lateral excitation.

A.2.3 Gain modulation

Recurrent networks with many–to–many connectivity (section A.1) show an effect known as gain modulation: adding a constant level to the input scales up the output curve, without modifying the tuning.

The simulated network shows is a similar effect (Fig. A.10(a)), but without a constant multiplicative factor (Fig. A.10(b)).

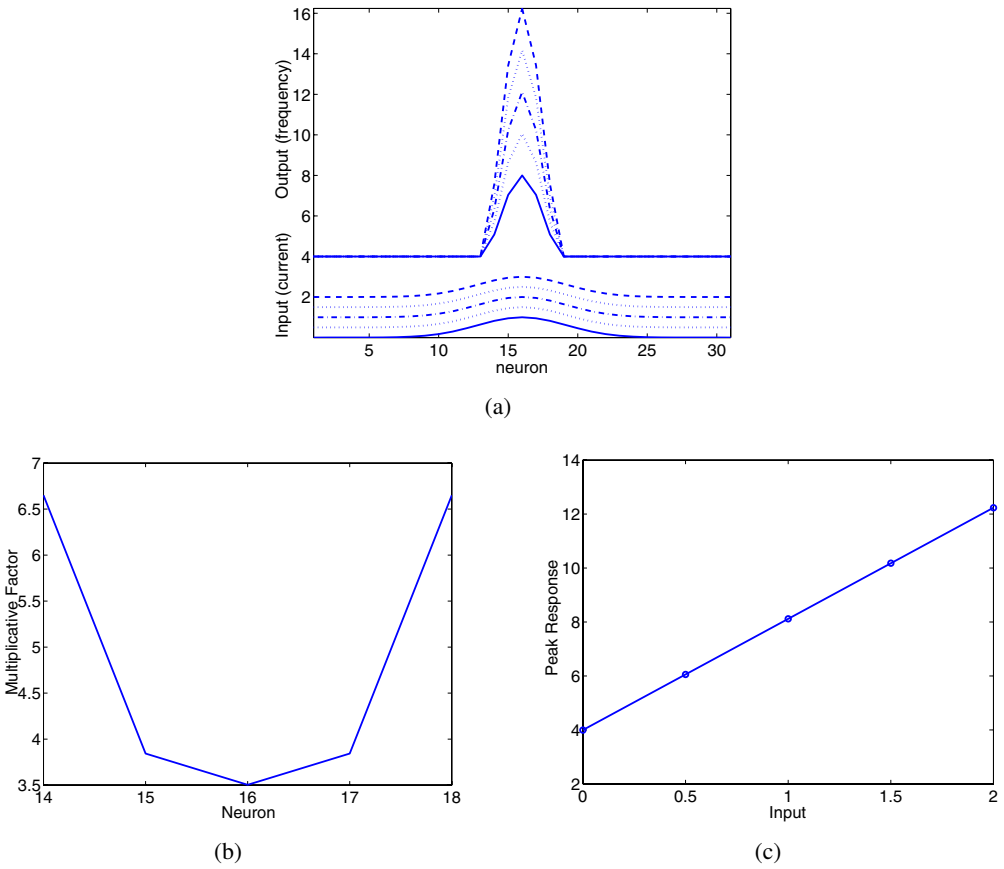


Figure A.10: Multiplicative effect for low coupling ($w_{ei} = 0.45$, $w_{ie} = 0.45$) and low lateral excitation ($w_{el} = 0.9$): (a) we add constant levels of current to the input, the network does not change its tuning, but the output scales up; (b) ratio between the output curve corresponding to "level 2" and the one corresponding to "level 0"; (c) the response of the most active neuron is plotted versus the constant level added to the input, resulting in a linear relationship

Appendix B

WTA circuit static and dynamic response properties

This appendix derives the static and dynamic response for the original WTA circuit proposed by Lazzaro et al. (1989) (Fig. 3.1 in the main text), and the equations for the diffusor network used in the chip for lateral connectivity, both excitatory, via transistors M_{exc}, and inhibitory, via transistors M_{inh} of Fig. 3.2.

B.1 Static response

Fig. B.1 shows the two cells WTA circuit and the corresponding small–signal model, describing the effect of small variations $i_{in,i}$ of the input current $I_{in,i}$ on the circuit voltages $V_{out,i}, V_{out,k}, V_c$, denoted by the small–signal voltages $v_{out,i}, v_{out,k}, v_c$, at a particular operating point $[I_{in,i}, I_{in,k}, I_{out,i}, I_{out,k}]$. The small–signal model of MOS transistors comprises a linear resistor r, in parallel with a linear dependent current source with conductance g:

$$g_{in,i} = \frac{kI_{in,i}}{U_T} \quad g_{out,i} = \frac{kI_{out,i}}{U_T} \quad r_{in,i} = \frac{V_e}{I_{in,i}} \quad r_{out,i} = \frac{V_e}{I_{out,i}}$$
$$g_{in,k} = \frac{kI_{in,k}}{U_T} \quad g_{out,k} = \frac{kI_{out,k}}{U_T} \quad r_{in,k} = \frac{V_e}{I_{in,k}} \quad r_{out,k} = \frac{V_e}{I_{out,k}}$$

where V_e is the Early voltage of the transistors, k is the subthreshold slope factor, a characteristic parameter of MOS transistors ranging from 0.6 to 0.8, and U_T is the thermal voltage. Applying Kirchhoff's current law to the subcircuits of Fig. B.1, and the (reasonable) approximation $V_e + \frac{U_T}{k} \cong V_e$, the circuit is described by a linear system of equations:

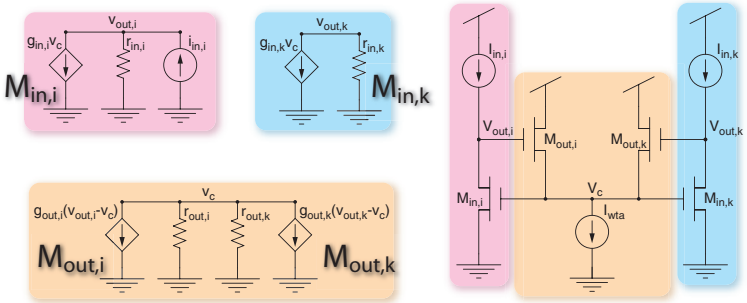


Figure B.1: Two cells WTA circuit and corresponding small signal model.

$$\frac{v_{out,i}}{i_{in,i}} = \frac{1}{I_{in,i}}\left(\frac{k}{U_T} + V_e\frac{I_{out,k}}{I_{wta}}\right) \tag{B.1}$$

$$\frac{v_{out,k}}{i_{in,i}} = V_e\frac{1}{I_{in,i}}\left(\frac{I_{out,i}}{I_{wta}}\right) \tag{B.2}$$

The first equation describes the winning cell i, the second describes the loosing cell k. The small–signal quantities in the equations can be expressed as differentials of the large–signal variables:

$$\frac{dV_{out,i}}{dI_{in,i}} = \frac{1}{I_{in,i}}\left(\frac{k}{U_T} + V_e\frac{I_{out,k}}{I_{wta}}\right) \tag{B.3}$$

$$\frac{dV_{out,k}}{dI_{in,i}} = V_e\frac{1}{I_{in,i}}\left(\frac{I_{out,i}}{I_{wta}}\right) \tag{B.4}$$

The output currents are described by the subthreshold transfer function of NMOS transistor in saturation region, describing transistors $M_{out,i}$ and $M_{out,k}$ respectively:

$$I_{out,i} = I_0 e^{\frac{kV_{out,i}-V_c}{U_T}} \tag{B.5}$$

$$I_{out,k} = I_0 e^{\frac{kV_{out,k}-V_c}{U_T}} \tag{B.6}$$

Applying again Kirchhoff's current law to the common node the biasing current I_{wta} can be expressed as the sum of the two output currents: $I_{wta} = I_{out,i} + I_{out,k}$. Substituting eq. (B.1) into this equation leads to:

$$I_{wta} = I_0 e^{\frac{kV_{out,i}-V_c}{U_T}} + I_0 e^{\frac{kV_{out,k}-V_c}{U_T}} \tag{B.7}$$

Dividing eq. (B.1) by eq. (B.7):

$$\frac{I_{out,i}}{I_{wta}} = \frac{1}{1 + e^{\frac{kV_{out,k}-V_{out,i}}{U_T}}} \tag{B.8}$$

$$\frac{I_{out,k}}{I_{wta}} = \frac{1}{1 + e^{\frac{kV_{out,i}-V_{out,k}}{U_T}}} \tag{B.9}$$

$$\tag{B.10}$$

Substituting this equations in eq. (B.1):

$$\frac{dV_{out,i}}{dI_{in,i}} = \frac{1}{I_{in,i}}\left(\frac{U_T}{k} + V_e\frac{1}{1 + e^{\frac{kV_{out,i}-V_{out,k}}{U_T}}}\right) \tag{B.11}$$

$$\frac{dV_{out,k}}{dI_{in,i}} = V_e\frac{1}{I_{in,i}}\left(\frac{1}{1 + e^{\frac{kV_{out,k}-V_{out,i}}{U_T}}}\right) \tag{B.12}$$

The behavior of the WTA circuit is described by a set of differential equations involving only $V_{out,i}$, $V_{out,k}$, and $I_{in,i}$. The voltages are expressed only in the subexpressions derived from eq. (B.1), that are *Fermi functions* of the difference $kV_{out,k} - V_{out,i}$, and $kV_{out,i} - V_{out,k}$, respectively. For $kV_{out,i} - V_{out,k} \gg U_T$ subexpression (B.12) is approximately zero,

while subexpression (B.12) is approximately one; viceversa for $kV_{out,k} - V_{out,i} \gg U_T$ subexpression (B.12) is approximately one, while subexpression (B.12) is approximately zero. In the crossover region it is reasonable to assume that the output voltages are changing with the same slope relative to the input current $I_{in,i}$; this assumption can be expressed with the approximations $kV_{out,i} - V_{out,k} \cong 2(kV_{out,i} - V_m)$, and $kV_{out,k} - V_{out,i} \cong 2(kV_{out,k} - V_m)$, where from the qualitative analysis in the main text $V_m = V_{out,i} = V_{out,k}$, when the input currents are equal. The simplified differential equations system becomes:

$$\frac{dV_{out,i}}{dI_{in,i}} = \frac{1}{I_{in,i}} \left(\frac{k}{U_T} + V_e \frac{1}{1 + e^{\frac{2(kV_{out,i} - V_m)}{U_T}}} \right) \tag{B.13}$$

$$\frac{dV_{out,k}}{dI_{in,i}} = V_e \frac{1}{I_{in,i}} \left(\frac{1}{1 + e^{\frac{2(kV_{out,k} - V_m)}{U_T}}} \right) \tag{B.14}$$

Straightforward integration of this equations, with the approximation $V_e + \frac{U_T}{k} \approx V_e$, leads to the closed–form approximation of the circuit response:

$$\ln \left(\frac{I_{in,i}}{I_m} \right) = \frac{V_{out,i} - V_m}{V_e} + \frac{1}{2} \ln \left(1 + \frac{U_T}{kV_e} e^{2\frac{kV_{out,i} - V_m}{U_T}} \right) \tag{B.15}$$

$$\ln \left(\frac{I_{in,i}}{I_m} \right) = \frac{V_m - V_{out,k}}{V_e} + \frac{1}{2} \frac{U_T}{kV_e} \left(1 - e^{2\frac{kV_{out,k} - V_m}{U_T}} \right) \tag{B.16}$$

These equations can be evaluated in three region of interest to derive an explicit approximation of the circuit response:

- $V_{out,i} \cong V_{out,k} \cong V_m$, i.e. when $I_{in,i} \to I_m$, then the closed-form solution can be linearized, yielding to the simpler relations:

$$V_{out,i} = \frac{V_e}{2} \left(\frac{I_{in,i}}{I_m} - 1 \right) + V_m \tag{B.17}$$

$$V_{out,k} = \frac{V_e}{2} \left(1 - \frac{I_{in,i}}{I_m} \right) + V_m \tag{B.18}$$

  both outputs are a linear function of the input current, with a slope of $\pm \frac{V_e}{2I_m}$. As described in the intuitive explanation of the circuit behavior in the main text, when the input currents are similar, the output voltages start moving linearly, decreasing (and increasing) the current in transistor $M_{in,k}$ ($M_{in,i}$) thanks to the Early effect.

- $V_{out,i} \gg V_m, V_{out,k} \ll V_m$, i.e. when $I_{in,i} = I_m + \delta$; in such a case eq. (B.1) can be simplified:

$$V_{out,i} = \frac{U_T}{k} \ln \frac{I_m + \delta}{I_m} + \frac{U_T}{2k} \ln \frac{kV_e}{U_T} + V_m \tag{B.19}$$

$$V_{out,k} = \frac{U_T}{2k} + V_m - V_e \ln \frac{I_m + \delta}{I_m} \tag{B.20}$$

  Eq. (B.1) holds with the approximation $\frac{U_T}{kV_e} e^{\frac{2(kV_{out,i} - V_m)}{U_T}} \gg 1$, that rearranged in $V_{out,i} - V_m / k \gg \frac{U_T}{2k} \ln \frac{kV_e}{U_T}$. That is, if k = 1, $V_{out,i} - V_m$ must be greater than

150mV, in typical fabrication processes. This error stems from the main approximation $kV_{out,i} - V_{out,k} \cong 2(kV_{out,i} - V_m)$, valid only if $kV_{out,i} - V_{out,k} \leq \frac{U_T}{k}$. Therefore the equation derived in the main text better predicts circuits winning behavior. The circuit's behavior is then approximated by the following equations:

$$V_{out,i} = \frac{U_T}{k^2} \ln \frac{I_m + \delta}{I_0} + \frac{U_T}{k} \ln \frac{I_{wta}}{I_0} \tag{B.21}$$

$$V_{out,k} = \frac{U_T}{2k} + V_m - \frac{V_e}{I_m}\delta \tag{B.22}$$

- $V_{out,i} \ll V_m, V_{out,k} \gg V_m$, i.e. $I_{in,i} < I_m$. Following the same procedure as above, the behavior of the circuit is:

$$V_{out,i} = V_m - \frac{V_e}{I_m}\delta \tag{B.23}$$

$$\ln\left(\frac{I_{in,i}}{I_m}\right) = \frac{V_m - V_{out,k}}{V_e} - \frac{1}{2}\frac{U_T}{kV_e} e^{2\frac{kV_{out,k} - V_m}{U_T}} \rightarrow -\infty \tag{B.24}$$

The losing responses of $V_{out,i}$ and $V_{out,k}$ are identical, as expected from the symmetry of the circuit; the winning response forces $I_{in,i} \rightarrow 0$, as above the problem stems from the approximation $kV_{out,i} - V_{out,k} \cong 2(kV_{out,i} - V_m)$, the equation derived in the main text better predicts the winning response:

$$V_{out,i} = V_m - \frac{V_e}{I_m}\delta \tag{B.25}$$

$$V_{out,k} = \frac{U_T}{k^2} \ln \frac{I_m + \delta}{I_0} + \frac{U_T}{k} \ln \frac{I_{wta}}{I_0} \tag{B.26}$$

In summary, the output voltage corresponding to the highest input current encodes logarithmically the corresponding input current, the voltage corresponding to the losing cells decreases with the separation between the input currents, with a slope set by the mean level of the input currents and the Early voltage. Increasing the Early voltage, with longer transistors, increases the resolution of the circuit, by narrowing its losing response.

B.2 Dynamic response

The dynamic behavior of the circuit can be modeled adding capacitors to the small–signal model of the previous section. Fig. B.2 shows the extended model, where the definition of resistors and conductances correspond to those listed in Sec. B.1, the capacitors added correspond to parasitic capacitances intrinsic to the circuit.

The solution of the resulting linear system is a function of the unknown large signals $I_{out,i}$ and $I_{out,k}$. However for the input conditions $I_{in,i} = I_m + \delta$, $I_{in,k} = I_m$, it is reasonable to approximate $I_{out,i} \approx I_{wta}$, and $I_{out,k} \approx 0$, even for relatively small values of δ, given the exponential dependence of these currents on the output voltages $V_{out,i}$ and $V_{out,k}$. With these

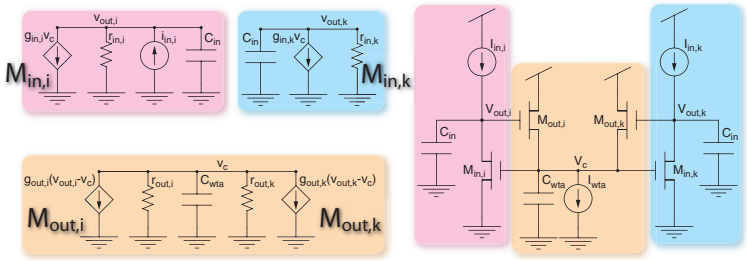


Figure B.2: Two cells WTA circuit and corresponding small signal model, with explicit capacitors added to model the dynamic response of the WTA.

approximations the ratio between the small-signal output voltages and the small-signal input current are:

$$\frac{v_{out,i}}{i_{in,i}} = \frac{U_T}{kI_{in,i}} \frac{\frac{C_{wta}U_T}{kI_{wta}}s+1}{\left(\frac{s}{a+b}+1\right)\left(\frac{s}{a-b}+1\right)} \tag{B.27}$$

$$\frac{v_{out,k}}{i_{in,i}} = -\frac{V_e}{I_{in,i}} \frac{1}{\left(\frac{C_{in}V_e}{I_{in,k}}s+1\right)\left(\frac{s}{a+b}+1\right)\left(\frac{s}{a-b}+1\right)} \tag{B.28}$$

where

$$a = \frac{I_{in,i}}{2C_{in}V_e} + \frac{kI_{wta}}{2C_cU_T} \tag{B.29}$$

$$b = \sqrt{\left(\frac{I_{in,i}}{2C_{in}V_e}\right)^2 + \left(\frac{kI_{wta}}{2C_cU_T}\right)^2 - \frac{k^2I_{wta}I_{in,i}}{C_{in}C_{wta}U_T^2}} \tag{B.30}$$

The system is stable if b has real poles, i.e. if $I_{wta} > 4I_{in,i}\frac{C_{in}}{C_{wta}}$. In such a case the system exhibits first order behavior and the first-order time constants for $V_{out,i}$ and $V_{out,k}$ are respectively $\tau_w = \frac{C_{in}U_T}{kI_m}$ and $\tau_l = \frac{C_{in}V_e}{I_m}$. The time constants depend on the Early voltage and on the mean level of the input currents. In the SAC chip an explicit capacitor is connected to the input node, to relax the stability condition derived with the dynamic analysis of the circuit.

B.3 Diffusor network

In his work Lazzaro et al. (1989) suggested a modification of the original WTA circuit that allows localized competition, it was implemented by substituting the direct connections between the current conveyors, with non–linear resistors. Unfortunately in CMOS technology it is not easily possible to implement big resistors; if only considering the diffusion of currents, diffusor networks can be implemented, replacing resistors with transistors with fixed gate voltage. Such networks are extensively used in silicon retinas (Choi et al., 2004; Liu and Boahen, 1996), since they produce center–surround spatial response. DeWeerth and Morris (1995) included a diffusor network to distribute hysteresis, Indiveri (2001a) used a diffusor network for implementing local competition, as suggested in (Lazzaro et al., 1989), and one to implement local excitation. He formally derived the diffusion equations for the local excitatory network that I report in this Appendix.

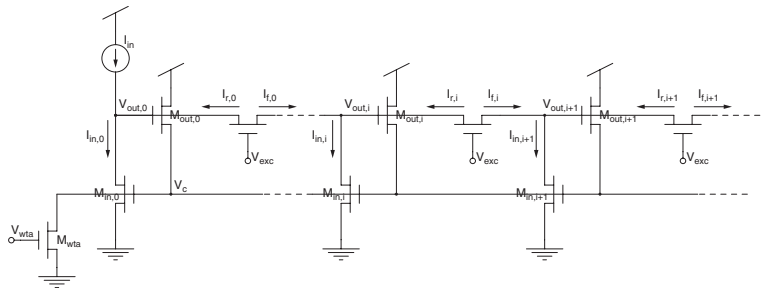


Figure B.3: Two cells WTA circuit and corresponding small signal model.

Fig. B.3 shows the simplified WTA network, without hysteresis, local competition, and source–degeneration, where an input current $I_{in} > 0$ is applied only to the first WTA node. In the subthreshold region of operation the current flowing though a single transistor can be divided into forward and reverse component, in the diffusors $I_{d,i} = I_{f,i} - I_{r,i}$, where

$$I_{f,i} = I_0 e^{\frac{kV_{exc}}{U_T} - \frac{V_{i+1}}{U_T}} \tag{B.31}$$

$$I_{r,i} = I_0 e^{\frac{kV_{exc}}{U_T} - \frac{V_i}{U_T}} \tag{B.32}$$

From these equations follows $I_{f,i} = I_{r,i+1}$. Kirchhoff's current law applied at each node i yields to:

$$I_{in,i} = (I_{f,i-1} - I_{r,i-1}) - (I_{f,i} - I_{r,i}) \tag{B.33}$$

merging these equations leads to the expression of the input current to the $i - th$ node in terms of reverse current in neighboring nodes $i - 1$ and $i + 1$:

$$I_{in,i} = 2I_{r,i} - I_{r,i-1} - I_{r,i+1} \tag{B.34}$$

The subthreshold transfer function of the transistor $M_{in,i}$ gives another formulation for the input current:

$$I_{in,i} = I_0 e^{\frac{kV_c}{U_T}} \left(1 - e^{-\frac{V_{out,i}}{U_T}}\right) \tag{B.35}$$

expressing $V_{out,i}$ in terms of $I_{r,i}$ using eq. (B.3)

$$I_{in,i} = I_0 e^{\frac{kV_c}{U_T}} - e^{k\left(\frac{V_c}{U_T} - \frac{V_{exc}}{U_T}\right)} I_{r,i} \tag{B.36}$$

from which follows

$$I_{r,i} = \lambda I_0 e^{\frac{kV_c}{U_T}} - \lambda I_{in,i} \tag{B.37}$$

where $\lambda = e^{-k\left(\frac{V_c}{U_T} - \frac{V_{exc}}{U_T}\right)}$. Substituting eq. (B.3) into eq. (B.3), yields to the discrete approximation of a Laplacian:

$$I_{in,i} = \lambda \left(I_{in,i-1} - 2I_{in,i} + I_{in,i+1}\right) \tag{B.38}$$

then:

$$I_{in,i} = \frac{\lambda}{1 + 2\lambda} I_{in,i-1} + \frac{\lambda}{1 + 2\lambda} I_{in,i+1} \tag{B.39}$$

using this equation recursively:

$$I_{in,i} = \frac{\lambda}{1+2\lambda} I_{in,i-1} + \frac{\lambda^2}{(1+2\lambda)^2}(I_{in,i} + I_{in,i+2}) \tag{B.40}$$

if $\lambda \ll 1$ this equation reduces to

$$I_{in,i} \approx \lambda I_{in,i-1} \tag{B.41}$$

The current flowing to ground through the input transistor of the $n-th$ WTA node can be estimated by recursively applying eq. (B.3) until reaching the node 0, where the input current is applied:

$$I_{in,n} \approx \lambda^n I_{in,0} \tag{B.42}$$

if $\lambda \ll 1$, $I_{in,0} \approx I_{in}$, the externally applied current, therefore:

$$I_{in,n} = I_{in} e^{-nk\left(\frac{V_c}{U_T} - \frac{V_{exc}}{U_T}\right)} \tag{B.43}$$

λ is the diffusor's space constant, it is exponentially dependent on $V_{exc} - V_c$; V_{exc} is an externally tunable bias, and V_c depends logarithmically on the input current I_{in}. In the simplified circuit of Fig. B.3

$$I_{in,0} = I_0 e^{\frac{kV_c}{U_T}} \approx I_{in} \tag{B.44}$$

From eq. (B.3) λ can be expressed in terms of V_{exc} and I_{in}, and eq. (B.3) reduces to:

$$I_{in,n} = I_{in} \left(\frac{I_0 e^{\frac{kV_exc}{U_T}}}{I_{in}} \right)^n \tag{B.45}$$

.

Appendix C

Integrating multiple AER and chip–control analysis tools

This appendix describes the setup used for the experiments in the main text. Figure C.1 shows the components of the system highlighting the division into hardware and software components used in the text. This appendix is meant to be a guide for setting up a similar system. The code included comprises the routines to merge and interconnect the different software modules developed for each part of the system.

C.1 hardware components

The hardware setup comprises the AER chips, the retina and the SAC described in the main text (see Sec. 5.3 and 4).

Retina

The silicon retina outputs its AER events using the SCX protocol. The addressing specification are defined by the "Caviar" standard; specifically the version used by the 64×64 version used in this work requires 6 bits for the X address, from b_1 to b_6, and 6 bits, from b_8 to b_{13} for the Y address, plus one bit (b_0) coding for the ON/OFF polarity of the event. Bits b_{14} and b_{15} are used to code for the chip identity (or AER channel). If the right bit is the least significant one the 16 bits of AER address are organized as follows:

AER_1 | AER_0 | $y_5\ y_4\ y_3\ y_2\ y_1\ y_0$ | ignored | $x_5\ x_4\ x_3\ x_2\ x_1\ x_0$ | (ON/OFF)

The chip's biases are controlled manually with potentiometers mounted on the chip's testing board.

SAC

The SAC receives AER events with P2P protocol, while using SCX for the output events. The 32×32 pixels space addressing requires 5 bits for the X and 5 bits for the Y address, consecutively mapped on the 16 AER address bits:

AER_1 | AER_0 | ignored (b_{13}/b_{10}) | $y_4\ y_3\ y_2\ y_1\ y_0$ | $x_4\ x_3\ x_2\ x_1\ x_0$

The same addressing specification is used for both input and output events.

Computer interfacing components

A Linux desktop is interfaced to the two chips.

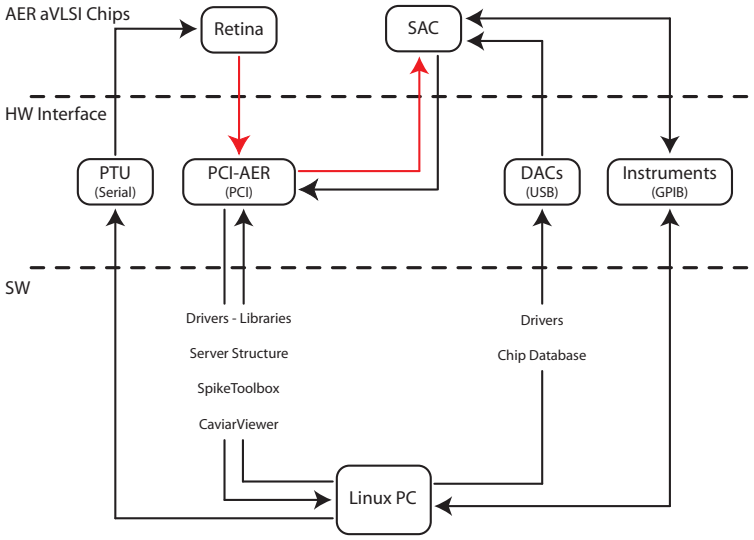


Figure C.1: Schematic diagram of the hardware and software components of the setup for an AER multi–chip system comprising a silicon retina and a post–processing chip such as the SAC. The system comprises the two chips, the interfacing devices to the host computer, and the software modules for the control of each device. The PCI–AER board serves for interfacing of any AER chip to the computer, and for the interconnection of chips. The libraries control the board access and settings, the server structure handles the interfacing to the board via the libraries, and allows the independent and parallel use of each board function. It is embedded in Matlab, and uses .mex files to handle the C–based libraries. The SpikeToolbox is used to generate, acquire, plot and analyze any spike train. The DAC boards set the specified voltage value on the chip; Matlab functions are available to access the drivers and a Chip Database allows the definition of the biases of each setup. The PTU orients the retina, it is controlled from Matlab, via serial port commands. The GPIB card interfaces the computer with the instruments, specifically, the function generator and the oscilloscope. Matlab functions and routines are available for the GPIB card functions.

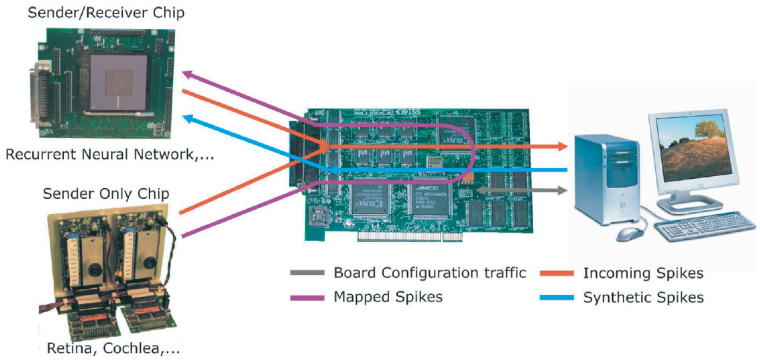


Figure C.2: Multi–chip AER communication framework based on the PCI–AER board. Colored arrows indicate the information flow for the different operating modes of the board. (adapted from (Dante et al., 2005))

A PCI–AER board (Dante and Del Giudice, 2001; Dante et al., 2005), developed in Rome at the "Istituto Superiore di Sanitá", connects the two chips to the computer, and routes the retina activity to the input of the SAC. Fig. C.2 shows the board and its functionality; it can handle multiple–chip AER–based systems, performing three main functions: monitoring, sequencing, and mapping. The monitor writes on a FIFO the events over the AER bus, adding a time–stamp (1μ s resolution) to each monitored address. The FIFO is then accessed via software and allows the visualization and analysis of the chips activity. The sequencer generates spike traffic on the AER bus, reading event lists generated by soft-

ware, emulating the activity of a chip connected to the PCI–AER board. The mapper maps incoming AER addresses to outgoing AER addresses via a look–up table. This function allows to implement any type of connectivity between different chips, and within the same chip. Examples of systems exploiting the functionality of the PCI–AER board are described in (Chicca et al., 2006b; Indiveri, 2005; Oster and Liu, 2005).

The parametric space of the SAC voltage biases can be explored thanks to a bank of Digital–to–Analog Converters (DAC), controlled via an USB interface, substituting the traditional potentiometers.

Some testing signals from the SAC can be connected to oscilloscope probes and the signal traces acquired via a GPIB card.

The system comprises also a Pan-Tilt Unit (PTU), controlled via a serial port, used to orient the center of the retina towards the target chosen by the SAC.

Finally a function generator, connected to the computer via the GPIB card, or an LCD screen can be used to generate stimuli for the retina.

C.2 software components

The possibility of connecting the chips to a computer via the AER system and the other interfaces listed in the previous section allows more extensive and automatized testing of our chips. A common effort of the hardware group in the Institute of Neuroinformatics has lead to the development of many software tools for the interfacing of the different elements comprised in the systems. As the systems increase in complexity and variety, a particular effort has been deployed to the design of general tools with wide applicability. My major role in this effort, besides β–testing for debugging, was to merge the different parts in a coherent frame, in particular I wrote some routines to link the different AER systems.

AER Software

The PCI–AER board is supported by softwares developed within the Institute of Neuroinformatics (INI); besides the driver and library functions created and maintained by A. Whatley (and E. Chicca), M. Oster (within the CAVIAR EU project framework) has developed a client–server architecture on top of the library to enable the use of the board on–line from within Matlab, including real–time data display (Oster et al., 2005).

On top of these utilities D.R. Muir (INI) has developed a Matlab toolbox, the SpikeToolbox, for the off–line generation and manipulation of spike trains to be sent to, or read from, the PCI–AER. I have merged both softwares for my experiments, to exploit the on–line visualization and continuous stimulation allowed by the client–server structure, with the sophisticated and rich spike handling offered by the SpikeToolbox; I also contributed to the SpikeToolbox adding few functions for analyzing bi–dimensional spike trains, and for the analysis of the scan–path.

Another tool for the visualization of the AER activity is the CaviarViewer, designed by T. Delbrück. It was originally designed to display the retina activity acquired via an USB–AER monitoring device; via an utility function *saveaerdat.m* it can also load Matlab data appropriately translated in the Caviar addressing format and structured in a two columns matrix with time–stamps and addresses. An example of function translating spike trains from the SpikeToolbox to the CaviarViewer .dat file format is *STAddrAERCaviarConvertSacRet64.m* shown below, where all of the retina events are saved as OFF events and all of the SAC events are saved as ON events, on the same addressing space. The CaviarViewer

will display the retina activity (in grey) and the SAC activity (in white) superimposed.

```
function stTrain = STAddrAERCaviarConvertSacRet64(stSpikeListRet,stSpikeListSac,stasRet,stasSac)

    % STAddrAERCaviarConvertSacRet64 - FUNCTION converts addresses of retina
    % 64x64 and SAC 32x32 to Caviar standard address specification. The retina
    % addresses are mapped to off events. The SAC addresses are mapped to on
    % events, and translated to 64x64 mapping
    %
    % Usage: [stTrain] =
    % STAddrAERCaviarConvertSacRet64(stSpikeListRet,stSpikeListSac,stasRet,stasSac)
    % stSpikeListRet: two columns matrix with [time_stamp addr_retina]
    % stSpikeListSac: two columns matrix with [time_stamp addr_sac]
    % stasRet: addressing specification of retina
    % stasSac: addressing specification of Sac

    % Author: Chiara Bartolozzi <chiara@ini.phys.ethz.ch>
    % Created: 8th Novemebr, 2006


    %-- Retina 64x64 --%
    [xAddr yAddr] = STAddrLogicalExtract(stSpikeListRet(:,2),stasRet);
    xAddr = bitshift(xAddr,1); % first addr of retina is ignored (on/off)
    xAddr = bitor(xAddr,1); % set all the events of retina with lsb at '1' => off events
    yAddr = bitshift(yAddr,8); % y addr starts from bit 8
    physAddrRet = bitor(xAddr,yAddr);
    %-- SAC --%
    [xAddr yAddr] = STAddrLogicalExtract(stSpikeListSac(:,2),stasSac);
    xAddr = bitshift(xAddr,2); % transform into 64x64, disregarding the 2
                               % lsb of addr and the 1st lsb on/off
    yAddr = bitshift(yAddr,9); % same as above
    physAddrSac = bitor(xAddr,yAddr);
    %-- create 2 column matrix with [time_stamp addr] --%
    stTrainRet = [stSpikeListRet(:,1) physAddrRet];
    stTrainSac = [stSpikeListSac(:,1) physAddrSac];
    %-- merge Retina and SAC trains --%
    stTrain = vertcat(stTrainRet,stTrainSac);
    stTrain = sortrows(stTrain,1);
end
%-- end of STAddrAERCaviarConvertSacRet64 --%
```

DAC interfacing software and Chip Database

Oster (2005) wrote a software that encapsulates in Matlab environment the functions to control the voltage value of each bias on the chip. The tool comprises a set of files for the specifications of the chip biases in terms of their name and function, their connections on the chip pins, and their connection to the DAC channels, these files are handled in a SubVersion (SVN) based chip database.

Putting everything together

A description on how to install the PCI–AER drivers, the DAC board driver, and the Chip Database can be found in the www.ini.uzh.ch *FAQ*: "How to start using DAC & PCI–AER". The SVN repositories needed are:
DAC board and Chip Database:

https://svn.ini.uzh.ch/repos/avlsi/CAVIAR/common/duckboard/DriverMatlab
https://svn.ini.uzh.ch/repos/avlsi/CAVIAR/common/duckboard/DriverLinux
https://svn.ini.uzh.ch/repos/avlsi/CAVIAR/common/duckboard/Documentation
https://svn.ini.uzh.ch/repos/avlsi/CAVIAR/common/ChipDatabase

Server Structure:
https://svn.ini.uzh.ch/repos/avlsi/CAVIAR/wp5/PCIAERg0/matlab
SpikeToolbox:
https://svn.ini.uzh.ch/repos/avlsi/common/SpikeToolbox
The script *ChipInitMap.m* comprises all of the initializations required by the PCI–AER, the SpikeToolbox, the server structure, and the DAC boards chip database, together with some examples of data visualization.

```
%------------------------------------------------------%
%--- CHIP INITIALIZATION FOR SAC AND RETINA ---%
%------------------------------------------------------%
% author: Chiara
clear all
%------------------------------------------------------%
%--- SPIKE TOOLBOX Initialising the AER parameters ---%
%------------------------------------------------------%
stOptions = STOptions;
%--- addressing specification of monitored PCIAER channels ---%
% ch0 -> monitor SAC
stOptions.MonitorChannelsAddressing{1} = STAddrSpecSynapse2DNeuron(0,5,5,0,31,31,0,0,0,0);
stasSAC = stOptions.MonitorChannelsAddressing{1};
% ch1 -> monitor retina
stOptions.MonitorChannelsAddressing{2} = STAddrSpecRetina64;
stasRetina =  STAddrSpecSynapse2DNeuron(0,6,6,0,63,63);
% ch2 -> not used
stOptions.MonitorChannelsAddressing{3} = [];
% ch3 -> not used
stOptions.MonitorChannelsAddressing{4} = [];
%--- addressing  specification of output PCIAER addresses (to the SAC) ---%
stOptions.stasDefaultOutputSpecification = STAddrSpecSynapse2DNeuron(0,5,5,0,31,31);
%--- Apply the new options configuration
STOptions(stOptions);
STDescribe(stOptions);
STOptionsSave;
%--- end of SpikeToolbox initialization ---%

%---------------------------------------%
%--- Server structure configuration ---%
%---------------------------------------%
%---  mount PCIAER board  ---%
%--- run only after reboot ---%
sudo /sbin/modprobe pciaer
%--- Server structure start ---%
%--- run only at the beginning of the experimental session ---%
str = pwd;
cd /local/d0/chiara/pciaer/matlab/
!./aerservers start
cmd = sprintf('cd␣%s',str)
eval(cmd)
%--- Server structure stop ---%
%--- run only at the end of experimental session ---%
% str = pwd;
% cd /local/d0/chiara/pciaer/matlab/
% !./aerservers stop % or "restart"
% cmd = sprintf('cd %s',str)
% eval(cmd)
%---  unmount PCIAER board  ---%
% sudo /sbin/modprobe -r pciaer
%--- PCIAER board config via servers structure ---%
%--- multi-chip system -> bit 15 and bit 14 are used for channel
% information
PciaerMapSetDemuxConfig;  % (2+14)
%--- input of SAC uses P2P protocol
PciaerMapSetProtocolP2P;
%--- MONITOR mask
% SAC -> ch0
```

```
% Retina -> ch1
MonSel = bin2dec('0011');
PciaerMonSetChannelSel(MonSel);
% -- MAPPER mask
% first it is disabled
MapSel = bin2dec('0000');
PciaerMapSetChannelSel(MapSel);
PciaerMapClearAllMappings;
% mapping four to one retina -> SAC
ch_ret = bin2dec('0100000000000000');% channel where the retina is attached (ch1)
ch_sac = bin2dec('1000000000000000');% channel where the input to SAC chip goes (ch2)
% address mapping 4 to 1
x = 0:63;
y = 0:63;
[rx ry] = meshgrid(x,y);
addrRet = STAddrPhysicalConstruct(stasRetina,rx,ry);
[sx sy] = meshgrid(floor(x./2),floor(y./2));
addrSAC = STAddrPhysicalConstruct(stasSAC,sx,sy);
% on events
set1to1 (n2addr(addrRet(:), 'retina64', 'on')+ch_ret, addrSAC(:)+ch_sac);
% off events
set1to1 (n2addr(addrRet(:), 'retina64', 'off')+ch_ret, addrSAC(:)+ch_sac);
% -- write look-up table and enable the mapping
PciaerMapSetOutputConfig1toMany;
MapSel = bin2dec('0010');
PciaerMapSetChannelSel(MapSel);
% -- when no mapping is used
% PciaerMapSetOutputConfigPassThru;
% -- functions to visualize the chips activity on-line during the
% experiment
% PciaerMonWatch('Mode','SpikeCount','Addresses',n2addr(0:4095, 'retina64', 'on')+ch_ret,'Dims',
% PciaerMonWatch('Mode','SpikeCount','Addresses',0:1023,'Dims',3)
% -- functions for data sequencing using stTrain generated with the
% SpikeToolbox
STStimServer(stTrain);
% -- functions for data acquisition
PciaerMonRecordEventsStart;
% wait for some time, the acquisition runs in background while
% matlab can do other functions
% then stop acquisition and load data in SpikeToolbox
% format, ch0 is for SAC, ch1 is for retina as specified above
[stMonSac stMonRet] = STMonServAcquire;
% otherwise, to record activity for tDuration
[stMonSac, stMonRet] = STMonServer(tDuration);
% -- end Servers and PCIAER board configuration -- %

% ------------------------------------ %
% -- DAC AMDA board configuration -- %
% ------------------------------------ %
% -- load saved bias settings (vValues)
load ../../ExpDef/WtaBiases
% -- get chip setup definition from chip database
getSetup('sac_chiara');
% -- get names of biases from DAC
cNames = getBiasNames;
% -- set all the biases to the loaded values
setAllBiases(cNames,vValues);
% -- end DAC configuration -- %
```

Pan–Tilt Unit interfacing software

The Pan–Tilt Unit is used to orient the retina. Three functions, *Fixation.m*, *Shake.m*, and *SaccadeTo.m*, are used to center the retina at fixation point $(0, 0)$, implement the micro–saccades around the current position, and saccade to a location, specified by the corresponding SAC address respectively.

```
function Fixation
% moves the pantilt unit to the fixation point (0,0)

% current angular position
global PhiX;
global PhiY;

if isempty (PhiX)
  disp ('initializing (x,y)=(0,0)');
  PhiX=0;
  PhiY=0;
  disp ('initializing pan-tilt unit');
  ptsend ('s')
end
ptsend('ps2000 ts2000 a')
cmdstring = sprintf('pp%d tp%d a', 0, 0);
ptsend(cmdstring)
% renew current position
PhiX = 0;
PhiY = 0;

function Shake
% -- micro sacceds: shakes the pan-tilt unit at the current position

global PhiX;
global PhiY;
if isempty (PhiX)
  disp ('initializing (x,y)=(0,0)');
  PhiX=0;
  PhiY=0;
  disp ('initializing pan-tilt unit');
  ptsend ('i');
  ptsend ('i pa1000 pu6000 ta1000 tu6000 ds');
end
% -- set speed for shaking
ptsend ('ps50 ts170 a');
cmdstring = sprintf ('m%d,%d,%d,%d', PhiX-1,PhiX+1,PhiY-1,PhiY+1);
ptsend (cmdstring);

function duration = SaccadeTo(AddrX,AddrY,chmaskflag)
% -- moves the pantilt unit to the given point (SAC address) -- %
% to work correctly, it needs the current position!
if nargin < 3
    chmaskflag = 1;
end

offset = 1.5;
vb = 57;
a  = 1000;
% -- Saccadic Suppression, disable any mapping -- %
PciaerMapSetChannelSel(0);
% -- stop PTU from micro-saccading -- %
ptstop;

% -- convert AddrX, AddrY in "angular steps" for the PTU -- %
% refer to center of the 2D array:
Center = 32/2;
AddrX = Center - AddrX; %
AddrY = AddrY - Center; %
dToLens =100; % distance in pixels from the retina to the lens
% angular position from "resting" PhiX = 0,PhiY = 0
AngleX = atand (AddrX / dToLens);
AngleY = atand (AddrY / dToLens);
% convert angle in degrees to PTU units (steps)
AngleX = AngleX / 185.1428 * 3600;
AngleY = AngleY / 185.1428 * 3600;
% round for PTU
AngleX = round(AngleX);
```

```
AngleY = round(AngleY);
% current angular position
global PhiX;
global PhiY;

if isempty (PhiX)
    Fixation;
    disp ('initializing (x,y)=(0,0)');
    PhiX=0;
    PhiY=0;
    disp ('initializing pan-tilt unit');
    ptsend ('s');
end
ptsend('ps2000 ts2000 a')
% AngleX and Y: angular distance respect to current position
% calculate absolute positions
newPhiX = (AngleX + PhiX);
newPhiY =  (AngleY + PhiY);
cmdstring = sprintf('pp%d tp%d a', newPhiX, newPhiY);
if (newPhiX == 0) && (newPhiY == 0)
    cmdstring = sprintf('pp%d tp%d a', PhiX, PhiY);
    return;
end
% -- send motion command to PTU via serial port -- %
ptsend(cmdstring)
% -- fixed duration of saccade (crude approximation) -- %
pause(.5)
% -- re-enable mapping, end of saccadic suppression -- %
PciaerMapSetChannelSel(2);
% -- renew current position
PhiX = newPhiX;
PhiY = newPhiY;
```

Chip stimulation and data acquisition from oscilloscope

The retina can be stimulated with any visual stimulus. In my experiments I have used also some synthetic stimuli created via software.

In Sec. 5.3.2 a function generator is used to light up a LED at a given frequency, and in Sec. 2 it is used to stimulate the DPI with voltage steps. The function generator is controlled from Matlab via a GPIB card. The code for the instrument control via the GPIB card can be found in */projects/avlsi/sw/linux-gpib/*.

Also the oscilloscope is connected via the GPIB card, that can be used both to acquire data and set the operation mode of the oscilloscope. These functionalities are extensively used in chapter 2 and 4.

Another mean for the stimulation of the retina is the LCD screen; the Matlab PsychToolbox, developed for psychophysics experiments, allows the design of any type of stimuli. The caveat for using this type of stimulation is that the frame based stimulation is not optimal for the retina, and can create artifacts. It can be useful for particular types of experiments as for example those requiring the rapid flashing of a stimulus (Chicca et al., 2006a). The PsychToolbox can be downloaded at *http://psychtoolbox.org/wikka.php?wakka=HomePage*.

The SaliencyToolbox was used to create saliency maps from static images. It can be downloaded at *http://www.saliencytoolbox.net/index.html*.

Abbreviations and Symbols

AER	Address Event Representation
$CAVIAR$	Convolution AER Vision Architecture for Real Time
CMI	Current Mirror Integrator
$CMOS$	Complementary Metal–Oxide–Semiconductor
DAC	Digital–to–Analog Converter
DPI	Diff–Pair Integrator
$EPSC$	Excitatory Post Synaptic Current
$EPSP$	Excitatory Post Synaptic Potential
FOA	Focus of attention
Gnd	Ground
$GPIB$	General Purpose Interfacing Board
$I\&F$	Integrate–and–Fire
IOR	Inhibition–of–Return
$IPSC$	Inhibitory Post Synaptic Current
$IPSP$	Inhibitory Post Synaptic Potential
ISI	Inter Spike Interval
LTU	Linear Threshold Unit
$MOSFET$	Metal–Oxide–Semiconductor Field Effect Transistor ($nFET$, $pFET$, $NMOS$,$PMOS$ abbreviations for n and p type transistors)
$NMDA$	N-methyl D-aspartate receptor
PC	Personal Computer
PCB	Printed Circuit Board
PCI	Peripheral Component Interconnect, a local bus standard developed by Intel Corporation
$P2P$	Point–to–Point AER protocol
PTU	Pan–Tilt Unit
SAC	Selective Attention Chip
SCX	Silicon Cortex AER protocol
STD	Short–Term Depression
$STDP$	Spike–Timing Dependent Plasticity
$VLSI$	Very Large Scale Integration
V_{dd}	Power supply
WTA	Winner–Take–All

Curriculum Vitae

Personal

Date of Birth	September 12^{th}, 1977
Nationality	Italian
Family status	Unmarried
Languages	Italian (mother language), English (fluent), French (beginner), German (beginner)

Education

Present	Doctorate of Sciences. Swiss Federal Institute of Technology (ETH), Institute of Neuroinformatics (INI), Zürich. Dissertation: *"Selective attention in silicon: from the design of an analog VLSI synapse to the implementation of a multi–chip system"*. Accepted on the reccomandation of: Prof. R. J. Douglas, Dr. G. Indiveri and Prof. P. König
November 2001	"Laurea" degree (Master equivalent) in Biomedical Engineering. Università degli Studi di Genova (Italy). Final mark: 110/110 summa cum laude. Degree Thesis in Perceptive Systems Models: *"Modelli architetturali di operatori corticali per la percezione binoculare del moto in profondità"* (Cortical operator's architectural models to binocularly perceive motion in depth) Advisor: Prof. Silvio Sabatini.
July 1996	High School Diploma. Liceo Scientifico "Leonardo da Vinci" – Genova (Italy). Final mark: 60/60

Teaching Experiences

Summer term 2006	Supervising semester project

Workshops

Jun 9^{th} – Jun 12^{th}, 2004	2^{nd} European School of Neuro–engineering "Massimo Grattarola"
Jun 29^{th} – Jul 19^{th}, 2003	Workshop of Neuromorphic Engineering - Telluride 2003

Fellowships

Mar 2003 - Mar 2004	*"Microelectronic devices for implementing neuromorphic models of selective attention"*. Supplied by CNR (Consiglio Nazionale delle Ricerche, Italy)
Feb 2002 – Apr 2002	*"Contribution for researchers training at high specialization research centers"*. Supplied by DIBE (Università di Genova), at the "Institute of Neuroinformatics" (ETH, Zürich)
Aug 1999 – Oct 1999	*"Erasmus"*. Supplied by the European Community.

Conferences

Nov 29^{th} - 31^{th}, 2006	IEEE Biomedical Circuits and Systems (BioCAS2006), London, UK
Oct 10^{th} - 12^{th}, 2006	Brain Inspired Cognitive Systems (BICS06), Lesvos Island, Greece
Aug 29^{th} - Sept 1^{st}, 2004	Brain Inspired Cognitive Systems (BICS04), Stirling, UK

VLSI design

Tanner Tools	Schematic capture (S-Edit), layout editor (L-Edit), LVS and simulation (T-Spice). L-Comp.

Computer Capabilities

Familiar with operating systems	Linux, Windows
Programming languages	Matlab

Publications

Journal Papers:

C. Bartolozzi and G. Indiveri **Synaptic dynamics in analog VLSI** , *Neural Computation* 2006 (in press)

C. Bartolozzi and G. Indiveri **Selective attention implemented with dynamic synapses and integrate-and-fire neurons**, *Neurocomputing*, Volume 69, Issues 16-18, October 2006, Pages 1971-1976, Brain Inspired Cognitive Systems

Refereed Conference Papers:

C. Bartolozzi and G. Indiveri **A spiking VLSI selective attention multi-chip system with dynamic synapses and integrate-and-fire neurons**, proceedings of *NIPS 2006*, December 2006, Vancouver, Canada

C. Bartolozzi, S. Mitra, G. Indiveri **An ultra low power current-mode filter with tunable gain for biomedical signal processing**, proceedings of *IEEE Biomedical Circuits and Systems (BioCAS 2006)*, Pages 130-133, November 2006, London, UK

C. Bartolozzi and G. Indiveri **Silicon synaptic homeostasis**, proceedings of *Brain Inspired Cognitive Systems (BICS06)*, October 2006, Lesvos Island, Greece

C. Bartolozzi and G. Indiveri **A neuromorphic selective attention architecture with dynamic synapses and Integrate-and-Fire neurons**, proceedings of *Brain Inspired Cognitive Systems (BICS04)*, August 2004, Stirling, UK

S. Sabatini, F. Solari, G. Andreani, C. Bartolozzi, G. M. Bisio **A hierarchical model of complex cells in visual cortex for the binocular perception of motion in depth**, proceedings of *NIPS 2001*, Dietterich, T.G. Becker, S. & Ghahramani, Z. (ed.) MIT Press, 2002, Vancouver, Canada

Bibliography

L.F. Abbott and W.G. Regehr. Synaptic computation. *Nature*, 431:796–803, October 2004.

L.F. Abbott, K. Sen, J. Varela, and S. Nelson. Synaptic depression and cortical gain control. *Science*, 275(5297):220–223, 1997.

AER. The address-event representation communication protocol AER 0.02. Caltech internal memo, February 1993. http://www.ini.unizh.ch/~amw/scx/std002.pdf.

M.S. Ambinder and D.J. Simons. *Neurobiology of attention*, chapter Attention capture: the interplay of expectations, attention, and awareness, pages 69–75. Elsevier academic press, 2005.

J. C. Anderson, T. Binzegger, K. A. C. Martin, and K. S. Rockland. The connection from cortical area v1 to v5: a light and electron microscopic study. *Journal of Neuroscience*, 18(24):10525–40, 1998.

J.S. Anderson, M. Carandini, and D. Ferster. Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *Journal of Physiology*, 84(2):909–926, August 2000.

J. Arthur and K. Boahen. Learning in silicon: Timing is everything. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006.

J.V. Arthur and K. Boahen. Recurrently connected silicon neurons with active dendrites for one-shot learning. In *IEEE International Joint Conference on Neural Networks*, volume 3, pages 1699–1704, July 2004.

C. Bartolozzi and G. Indiveri. Silicon synaptic homeostasis. In *Brain Inspired Cognitive Systems 2006*, pages 1–4, October 2006.

C. Bartolozzi and G. Indiveri. Synaptic dynamics in analog VLSI. *Neural Computation*, 2007. (In press).

R. Ben-Yishai, R. Lev Bar-Or, and H. Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences of the USA*, 92(9): 3844–3848, April 1995.

N.B. Bichot and J.D. Schall. *Neurobiology of attention*, chapter Prefrontal selection and control of covert and overt orienting, pages 117–123. Elsevier academic press, 2005.

K. A. Boahen. Point-to-point connectivity between neuromorphic chips using address-events. *IEEE Transactions on Circuits and Systems II*, 47(5):416–34, 2000.

K. A. Boahen. A burst-mode word-serial address-event link – I: Transmitter design. *IEEE Circuits and Systems I*, 51(7):1269–80, 2004.

K. A. Boahen. *Retinomorphic Vision Systems: Reverse Engineering the Vertebrate Retina*. Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1997.

K.A. Boahen. Neuromorphic microchips. *Scientific American*, pages 56–63, May 2005.

K.A. Boahen. Communicating neuronal ensembles between neuromorphic chips. In T. S. Lande, editor, *Neuromorphic Systems Engineering*, pages 229–259. Kluwer Academic, Norwell, MA, 1998.

M Boegerhausen, P Suter, and S.-C. Liu. Modeling short-term synaptic depression in silicon. *Neural Computation*, 15(2):331–348, Feb 2003.

A. Bofill, A.F. Murray, and D.P. Thompson. Circuits for VLSI implementation of temporally asymmetric hebbian learning. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

T.H. Borgstrom, M Ismail, and S.B. Bibyk. Programmable current-mode neural network for implementation in analogue MOS VLSI. *IEE Proceedings G*, 137(2):175–184, 1990.

V. Brajovic and T. Kanade. Computational sensor for visual tracking with attention. *IEEE Journal of Solid State Circuits*, 33(8):1199–1207, August 1998.

J. Burrone and V.N. Murthy. Synaptic gain control and homeostasis. *Current Opinion in Neurobiology*, 13:560–567, 2003.

M. Carandini, D. J. Heeger, and J. A. Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21): 8621–8644, 1997.

R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46:4333–4345, 2006. doi: 10.1016/j.visres.2006.08.019.

V. Chan, A. van Schaik, and S.-C. Liu. Spike response properties of an aer ear. In *Proceedings of IEEE International Symposium on Circuits and Systems*. IEEE, 2006.

F. S. Chance and L.F. Abbott. Input–specific adaptation in complex cells through synaptic depression. *Neurocomputing*, 38(40):141–46, 2001.

F. S. Chance, S. B. Nelson, and L. F. Abbott. Synaptic depression and the temporal response characteristics of V1 cells. *The Journal of Neuroscience*, 18(12):4785–99, 1998.

F.S. Chance, L.F. Abbott, and A.D. Reyes. Gain modulation from background synaptic input. *Neuron*, 35:773–782, August 2002.

E. Chicca. *A Neuromorphic VLSI System for Modeling Spike–Based Cooperative Competitive Neural Networks*. PhD thesis, ETH Zürich, Zürich, Switzerland, April 2006.

E. Chicca, D. Badoni, V. Dante, M. D'Andreagiovanni, G. Salina, S. Fusi, and P. Del Giudice. A VLSI recurrent network of integrate–and–fire neurons connected by plastic synapses with long term memory. *IEEE Transactions on Neural Networks*, 14(5):1297–1307, September 2003a.

E. Chicca, G. Indiveri, and R.J. Douglas. An adaptive silicon synapse. In *Proceedings of IEEE International Symposium on Circuits and Systems*, pages I–81–I–84. IEEE, May 2003b.

E. Chicca, P. Lichtsteiner, T. Delbrück, G. Indiveri, and R.J. Douglas. Modeling orientation selectivity using a neuromorphic multi-chip system. In *Proceedings of IEEE International Symposium on Circuits and Systems*, pages 1235–1238. IEEE, 2006a.

E. Chicca, A. M. Whatley, V. Dante, P. Lichtsteiner, T. Delbrück, P. Del Giudice, R. J. Douglas, and G. Indiveri. A multi-chip pulse-based neuromorphic infrastructure and its application to a model of orientation selectivity. *IEEE Transactions on Circuits and Systems I, Regular Papers*, 2006b. (In press).

J. Choi and B.J. Sheu. A high-precision VLSI winner-take-all circuit for self- organizing neural networks. *IEEE J. Solid-State Circuit*, 28(5):576–584, May 1993.

T. Y. W. Choi, B. E. Shi, and K. Boahen. An on-off orientation selective address event representation image transceiver chip. *IEEE Transactions on Circuits and Systems I*, 51(2):342–353, 2004.

C. L. Colby and M. E. Goldberg. Space and attention in parietal cortex. *Annu Rev Neurosci*, 22:319–49, 1999.

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

V. Dante, P. Del Giudice, and A. M. Whatley. PCI-AER – hardware and software for interfacing to address-event based neuromorphic systems. *The Neuromorphic Engineer*, 2(1):5–6, 2005. http://ine-web.org/research/newsletters/index.html.

S. Danziger, A. Kingstone, and J.J. Snyder. Inhibition of return to successively stimulated locations in a sequential visual search paradigm. *Journal of experimental psychology: Human perception and performance*, 24(5):1467–1475, 1998.

P. Dayan and L.F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001.

G. Deco and T.S. Lee. A unified model of spatial and object attention based on inter–cortical biased competition. *Neurocomputing*, 44–46:775–781, 2002.

S.R. Deiss, R.J. Douglas, and A.M. Whatley. A pulse-coded communications infrastructure for neuromorphic systems. In W. Maass and C. M. Bishop, editors, *Pulsed Neural Networks*, chapter 6, pages 157–78. MIT Press, 1998.

T. Delbrück. CaviarViewer application. http://www.ini.unizh.ch/ tobi/caviar/INI-AE-Biasgen/, November 2006.

T. Delbrück. Analog VLSI phototransduction by continuous-time, adaptive, logarithmic photoreceptor circuits. Technical report, California Institute of Technology, Pasadena, CA, 1994. CNS Memo No. 30.

T. Delbrück and C.A. Mead. Analog VLSI phototransduction by continuous-time, adaptive, logarithmic photoreceptor circuits. In C. Koch and H. Li, editors, *Vision Chips: Implementing vision algorithms with analog VLSI circuits*, pages 139–161. IEEE Computer Society Press, Los Alamitos, CA, 1995.

A. Demosthenous, S. Smedley, and J.erolla Taylor. A CMOS analog winner-take-all network for large-scale applications. *IEEE Transactions on Circuits and Systems I*, 45(3):300–304, March 1998.

N. S. Desai, R. H. Cudmore, S. B. Nelson, and G.G. Turrigiano. Critical periods for experience-dependent synaptic scaling in visual cortex. *Nature Neuroscience*, 5(8): 783–789, August 2002.

A. Destexhe, Z.F. Mainen, and T.J. Sejnowski. *Methods in Neuronal Modelling, from ions to networks*, chapter Kinetic Models of Synaptic Transmission, pages 1–25. The MIT Press, Cambridge, Massachussets, 1998.

S.P. DeWeerth and T.G Morris. CMOS current mode winner-take-all circuit with distributed hysteresis. *Electronics Letters*, 31(13):1051–1053, June 1995.

M.D. Dodd and J. Pratt. Rapid onset and long–term inhibition of return in the multiple cuing paradigm. *Psychological Research*, 2006. doi: DOI10.1007/s00426-006-0048-4.

R.J. Douglas, M.A. Mahowald, and K.A.C. Martin. Hybrid analog-digital architectures for neuromorphic systems. In *Proc. IEEE World Congress on Computational Intelligence*, volume 3, pages 1848–1853. IEEE, 1994.

R.J. Douglas, M.A. Mahowald, and C. Mead. Neuromorphic analogue VLSI. *Annu. Rev. Neurosci.*, 18:255–281, 1995.

R.J. Douglas, M.A. Mahowald, K.A.C. Martin, and K.J. Stratford. The role of synapses in cortical computation. *Journal of Neurocytology*, 25:893–911, 1996.

R.J. Douglas, C. Koch, M.A. Mahowald, and K.A.C. Martin. The role of recurrent excitation in neocortical circuits. In P Ulinski, editor, *Cerebral Cortex*. Plenum Press, 1999.

W. Einhäuser and P. König. Does luminance–contrast contribute to a saliency map for overt visual attention? *European journal of neuroscience*, 17:1089–1097, 2003. doi: 10.1046/j.1460-9568.2003.02508.x.

W. Einhäuser, W. Kruse, K.P. Hoffmann, and P. König. Differences of monkey and human overt attention under natural conditions. *Vision Research*, 46(8–9):1194–1209, Apr 2006.

V.P. Ferrera and S.G. Lisberger. Attention and target selection for smooth pursuit eye movements. *The Journal of Neuroscience*, 15(11):7472–7484, November 1995.

J.M. Findlay. *Neurobiology of attention*, chapter Covert attention and saccadic eye movements, pages 114–116. ElsevierAcademic press, 2005.

J.M. Findlay and R. Walker. A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and brain sciences*, 22(4):661–721, 1999.

A. Fish, A. Spivakovsky, A. Golberg, and O. Yadid-Pecht. VLSI sensor for multiple targets detection and tracking. *IEEE*, pages 543–546, 2004.

A. Fish, V. Milrud, and O. Yadid-Pecht. High-speed and high precision current winner-take-all circuit. *IEEE Transactions on circuits and systems-II: Express briefs*, 52(3):131–135, March 2005.

D. Frey. Future implications of the log domain paradigm. In *IEE Proceedings Circuits Devices Systems*, volume 147, pages 65–72, February 2000.

D. R. Frey. Log-domain filtering: An approach to current-mode filtering. *IEE Proceedings G: Circuits, Devices and Systems*, 140(6):406–416, December 1993.

S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D. J. Amit. Spike–driven synaptic plasticity: theory, simulation, VLSI implementation. *Neural Computation*, 12:2227–58, 2000.

A. Gara, M. A. Blumrich, D. Chen, G. L.-T. Chiu, P. Coteus, M. E. Giampapa, R. A. Haring, P. Heidelberger, D. Hoenicke, G. V. Kopcsay, T. A. Liebsch, M. Ohmacht, B. D. Steinmacher-Burow, T. Takken, and P. Vranas. Overview of the blue gene/l system architecture. *IBM Journal of research & development*, 49(2/3):195–212, March/May 2005.

J.L. Gardner and S.G. Lisberger. Serial linkage of target selection for orienting ad tracking eye movements. *Nature neuroscience*, 5:892–899, September 2002. doi: 10.1038/nn897.

C. Gordon, E. Farquhar, and P. Hasler. A family of floating-gate adapting synapses based upon transistor channel models. In *2004 IEEE International Symposium on Circuits and Systems*, volume 1, pages 317–20, May 2004.

M. Grattarola and G. Massobrio. *Bioelectronics handbook : MOSFETs, biosensors, and neurons*. McGraw-Hill, New York, 1998.

Z. S. Günay and E. Sánchez-Sinencio. CMOS winner-take-all circuits: a detail comparison. In *IEEE International Symposium on Circuits and Systems*, pages 41–44, June 1997.

R. Gütig and H. Sompolinsky. The tempotron: a neuron that learns spike timing–based decisions. *Nature Neuroscience*, 9:420–428, 2006. doi: 10.1038/nn1643.

R. Guyonneau, H. Kirchner, and S.J. Thorpe. Animals roll around the clock: The rotation invariance of ultrarapid visual processing. *Journal of Vision*, 6(10):1008–1017, 2006. doi: doi:10.1167/6.10.1. URL `http://journalofvision.org/6/10/1/`.

Ph. Häfliger and F. Bergh. An integrated circuit computing shift in stereo pictures using time domain spike signals. In *NORCHIP 2002*, 2002.

R. Hahnloser, R. Sarpeshkar, M.A. Mahowald, R.J. Douglas, and S. Seung. Digital selection and analog amplification co-exist in an electronic circuit inspired by neocortex. *Nature*, 405(6789):947–951, 2000.

J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Reading, MA, 1991.

A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117:500–44, 1952.

J. Hoffman and B. Subramaniam. The role of visual attention in saccadic eye movements. *Perception and Psychophysics*, 57(6):787–795, 1995.

T. Horiuchi and K. Hynna. A VLSI-based model of azimuthal echolocation in the big brown bat. *Autonomous Robots*, 11(3):241–247, 2001.

T. Horiuchi and E. Niebur. Conjunction search using a 1-D, analog VLSI-based attentional search/tracking chip. In D. Scott Wills and Stephen P. DeWeerth, editors, *Proc. Conf. Advanced Research in VLSI*, pages 276–290. IEEE Computer Society, 1999.

T. Horiuchi, T.G. Morris, C. Koch, and S.P. DeWeerth. Analog VLSI circuits for attention-based, visual tracking. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 706–712. MIT Press, 1997.

T. K. Horiuchi and C. Koch. Analog VLSI-based modeling of the primate oculomotor system. *Neural Computation*, 11(1):243–265, January 1999.

T.S. Horowitz and J.M. Wolfe. Visual search has no memory. *Nature*, 394(6693): 575–577, August 1998. doi: 10.1038/29068. URL `http://dx.doi.org/10.1038/29068`.

T.S. Horowitz, A.O. Holcombe, J.M. Wolfe, H.C. Arsenio, and J.S. DiMase. Attentional pursuit is faster than attentional saccade. *journal of vision*, 4:585–603, July 2004. doi: 10.1167/4.7.6.

D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Jour. Physiology*, 160:106–54, 1962.

D.H. Hubel and T.N. Wiesel. The functional architecture of the macaque visual cortex. the ferrier lecture. In *Proc. R. Soc. Lond. B*, volume 198, pages 1–59, 1977.

K. Hynna and K. Boahen. Space–rate coding in an adaptive silicon neuron. *Neural Networks*, 14:645–656, 2001.

K. M. Hynna and K. Boahen. Neuronal ion-channel dynamics in silicon. In *2006 IEEE International Symposium on Circuits and Systems*, pages 3614 – 3617, May 2006.

A.E. Iapata, A.L. Gee, J. Gottlieb, J.W. Biskley, and M.E. Goldberg. Lip responses to a pop—out stimulus are reduced if it is overtly ignored. *Nature Neuroscience*, 9 (8):1071–1076, August 2006. doi: 10.1038/nn1734.

G. Indiveri. Modeling selective attention using a neuromorphic analog VLSI device. *Neural Computation*, 12(12):2857–2880, December 2000a.

G. Indiveri. A 2D neuromorphic VLSI architecture for modeling selective attention. In S.-I. Amari, C. L. Giles, M. Gori, and V. Piuri, editors, *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks; IJCNN2000*, volume IV, pages 208–213. IEEE Computer Society, 2000b.

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

G. Indiveri. A neuromorphic VLSI device for implementing 2-D selective attention systems. *IEEE Transactions on Neural Networks*, 12(6):1455–1463, November 2001b.

G. Indiveri. Neuromorphic bistable VLSI synapses with spike-timing-dependent plasticity. In *Advances in Neural Information Processing Systems*, volume 15, pages 1091–1098, Cambridge, MA, December 2002. MIT Press.

G. Indiveri. Neuromorphic selective attention systems. In *Proc. IEEE International Symposium on Circuits and Systems*, pages III–770–III–773. IEEE, May 2003a.

G. Indiveri. A low-power adaptive integrate-and-fire neuron circuit. In *Proc. IEEE International Symposium on Circuits and Systems*, pages IV–820–IV–823. IEEE, May 2003b.

G. Indiveri. VLSI reconfigurable networks of integrate-and-fire neurons with spike-timing dependent plasticity. *The Neuromorphic Engineer*, 2(1):4–7, 2005. http://ine-web.org/research/newsletters/index.html.

G. Indiveri. Neuromorphic analog VLSI sensor for visual tracking: Circuits and application examples. *IEEE Transactions on Circuits and Systems II*, 46(11):1337–1347, November 1999.

G. Indiveri, R. Mürer, and J. Kramer. Active vision using an analog VLSI model of selective attention. *IEEE Transactions on Circuits and Systems II*, 48(5):492–500, May 2001.

G. Indiveri, P. Oswald, and J. Kramer. An adaptive visual tracking sensor with a hysteretic winner-take-all network. In *Proc. IEEE International Symposium on Circuits and Systems*, pages 324–327. IEEE, May 2002.

G. Indiveri, E. Chicca, and R. Douglas. A VLSI array of low-power spiking neurons and bistable synapses with spike–timing dependent plasticity. *IEEE Transactions on Neural Networks*, 17(1):211–221, Jan 2006.

L. Itti. Real–time high–performance attention focusing in outdoors color video streams. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging VII*, volume 4662, pages 235–243. SPIE, 2002. URL `http://link.aip.org/link/?PSI/4662/235/1`.

L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304– 1318, October 2004.

L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, Aug 2005a.

L. Itti. Quantitative modeling of perceptual salience at human eye position. *Visual Cognition*, 2005b.

L. Itti and C. Koch. A saliency–based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, May 2000.

L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

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

Laurent Itti, Carl Gold, and Christof Koch. Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9):1784–1793, 2001. URL `http://link.aip.org/link/?JOE/40/1784/1`.

B. Jagadeesh, L. Chelazzi, M. Mishkin, and R. Desimone. Learning increases stimulus salience in anterior inferior temporal cortex of the macaque. *journal of neurophysiology*, 86:290–303, 2001.

E. R. Kandel, J.H. Schwartz, and T. M. Jessell. *Principles of Neural Science*. Mc Graw Hill, 2000.

S. Kastner and L. G. Ungerleider. Mechanisms of visual attention in the human cortex. *Annu Rev Neurosci*, 23:315–41, 2000.

B. Khurana and E. Kowler. Shared attentional control of smooth eye movement and perception. *Vision Research*, 27(9):1603–1618, 1987. doi: doi:10.1016/0042-6989(87)90168-4.

D.C. Kiper and M. Carandini. *The neural basis of pattern vision*. Macmillan, 2003. in press.

R.M. Klein. Inhibition of return. *Trends in cognitive sciences*, 4(4):138–147, April 2000.

C. Koch. *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, 1999a.

C. Koch. *Biophysics of Computation: Information Processing in Single Neurons*, chapter Synaptic input, pages 85–116. Oxford University Press, 1999b.

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

C. Koch and D. Walther. Bottom–up visual attention to salient proto–object regions. In *Annual Meeting of the Vision Sciences Society*, Sarasota, F, 2006.

C. Koch, T. Poggio, and V. Torre. Nonlinear interactions in a dendritic tree: Localization, timing, and role in information processing. *PNAS*, 80:2799–2802, May 1983.

S.B. Laughlin and T.J. Sejnowski. Communication in neuronal networks. *Science*, 301(5641):1870–1874, 2003. doi: 10.1126/science.1089662. URL `http://www.sciencemag.org/cgi/content/abstract/301/5641/1870`.

J. Lazzaro, S. Ryckebusch, M.A. Mahowald, and C.A. Mead. Winner-take-all networks of $O(n)$ complexity. In D.S. Touretzky, editor, *Advances in neural information processing systems*, volume 2, pages 703–711, San Mateo - CA, 1989. Morgan Kaufmann.

J. Lazzaro, J. Wawrzynek, M. Mahowald, M. Sivilotti, and D. Gillespie. Silicon auditory processors as computer peripherals. *IEEE Transactions on Neural Networks*, 4:523–528, 1993.

J. P. Lazzaro. *Silicon Implementation of Pulse Coded Neural Networks*, chapter Low-power silicon axons, neurons, and synapses, pages 153–164. Kluwer Academic Publishers, 1994.

D.K. Lee, L. Itti, C. Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, 2(4):375–381, April 1999. doi: 10.1038/7286. URL `http://dx.doi.org/10.1038/7286`.

P. Lichtsteiner and T. Delbrück. A 64x64 aer logarithmic temporal derivative silicon retina. In *Research in Microelectronics and Electronics, 2005 PhD*, volume 2, pages 202–205, July 2005.

P. Lichtsteiner, T. Delbrück, and J. Kramer. Improved ON/OFF temporaly differentiating address-event imager. In *11th IEEE International Conference on Electronics, Circuits and Systems.*, pages 211–214. IEEE, December 2004.

P. Lichtsteiner, T. Delbrück, and C. Posch. A 100db dynamic range high-speed dual-line optical transient sensor with asynchronous readout. In *Proceedings of IEEE International Symposium on Circuits and Systems*. IEEE, 2006a.

P. Lichtsteiner, C. Posch, and T. Delbrück. A 128×128 120dB 30mW asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE ISSCC Digest of Technical Papers*, pages 508–509. IEEE, 2006b.

S. Liu and K.A. Boahen. Adaptive retina with center-surround receptive field. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in neural information processing systems*, volume 8. MIT Press, 1996.

S.-C. Liu, J. Kramer, G. Indiveri, T. Delbrück, T. Burg, and R. Douglas. Orientation-selective aVLSI spiking neurons. *Neural Networks*, 14(6/7):629–643, 2001.

S.-C. Liu, J. Kramer, G. Indiveri, T. Delbrück, and R. Douglas. *Analog VLSI:Circuits and Principles*. MIT Press, 2002.

S. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of neurophysiology*, 77:24–42, 1997.

W. Maass. On the computational power of winner-take-all. *Neural Computation*, 2000.

M.A.C. Maher, S.P. DeWeerth, M.A. Mahowald, and C.A. Mead. Implementing neural architectures using analog VLSI circuits. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS*, 36(5):643–652, May 1989.

M. Mahowald and R. Douglas. A silicon neuron. *Nature*, 354:515–518, 1991.

M.A. Mahowald. Analog VLSI chip for stereocorrespondence. In *IEEE International Symposium on Circuits and Systems, 1994. ISCAS94*, volume 6, pages 347–350, May 1994.

K. A. C. Martin. The microcircuits of visual cortex. *Current Opinion in Neurobiology*, 12:418–425, 2002.

S. Martinez-Conde, S.L. Macknik, and D.H. Hubel. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5(3):229–240, March 2004. doi: 10.1038/nrn1348.

S. Martinez-Conde, S.L. Macknik, X.G. Troncoso, and T.A. Dyar. Microsaccades counteract visual fading during fixation. *Neuron*, 49(2):297–305, January 2006.

J.C. Martinez-Trujillo and S. Treue. Feature–based attention increases the selectivity of population responses in primate visual cortex. *Current biology*, 14:744–751, May 2004. doi: 10.1016/j.cub.2004.04.028.

C.J. McAdams and J.H.R. Maunsell. Attention to both space and feature modulates neuronal responses in macaque area v4. *Journal of Neurophysiology*, 83:1751–1755, 2000.

K. McDermott, J.B. Mulligan, G. Bebis, and M.A. Webster. Visual search and eye movements in novel and familiar contexts. In B.E. Rogowitz, T.N. Pappas, and S.J. Daly, editors, *Human Vision and Electronic Imaging XI*, volume 6057, pages 0C–1–0C–12. SPIE, 2006.

C. Mead. Neuromorphic electronic systems. *Proceedings of the IEEE*, 78(10):1629–36, October 1990.

C. A. Mead, X. Arreguit, and J. Lazzaro. Analog VLSI model of binaural hearing. *IEEE Transactions on Neural Networks*, 2(2):230–236, March 1991.

C.A. Mead. *Analog VLSI and Neural Systems*. Addison-Wesley, Reading, MA, 1989.

P. Merolla and K. Boahen. A recurrent model of orientation maps with simple and complex cells. In *Advances in Neural Information Processing Systems*, volume 16, pages 995–1002. MIT Press, December 2004.

S. Mitra, S. Fusi, and G. Indiveri. A VLSI spike-driven dynamic synapse which learns. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 2777–2780. IEEE, May 2006.

H.M. Morgan, M.A. Paul, and S.P. Tipper. Inhibition of return can be associated with object identity but not to object category. *European journal of cognitive psychology*, 17(4):499–520, 2005. doi: 10.1080/09541440440000186.

R.G.M. Morris, S. Davis, and S.P. Butcher. Hippocampal synaptic plasticity and NMDA receptors: a role in information storage? *Philosophical Transactions: Biological Sciences*, 329(1253):187–204, August 1990.

T.G. Morris, C.S. Wilson, and S.P. DeWeerth. Analog VLSI circuits for sensory attentive processing. In *IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Dec 1996.

A. F. Murray. Pulse-based computation in VLSI neural networks. In W. Maass and C. M. Bishop, editors, *Pulsed Neural Networks*, chapter 3, pages 87–109. MIT Press, 1998.

A. F. Murray, D. Del Corso, and L. Tarassenko. Pulse-stream VLSI neural networks mixing analog and digital techniques. *IEEE Transactions on neural networks*, 2(2): 193–204, March 1991.

D. P. M. Northmore and J. G. Elias. Building silicon nervous systems with dendritic tree neuromorphs. In W. Maass and C. M. Bishop, editors, *Pulsed Neural Networks*, chapter 5, pages 135–156. MIT Press, 1998.

I. Ohzawa, G. Sclar, and Freeman R.D. Contrast gain control in the cat's visual system. *Journal of Neurophysiology*, 54(3):651–667, 1985.

M. Oster. Tuning aVLSI chips with a mouse click. *The Neuromorphic Engineer*, 2 (1):9, 2005. http://ine-web.org/research/newsletters/index.html.

M Oster and S.-C. Liu. Spiking inputs to a winner–take–all network. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 18, pages 1051–1058, Cambridge, MA, December 2005. Neural Information Processing Systems Foundation, MIT Press.

M. Oster, A. M. Whatley, S.-C. Liu, and R. J. Douglas. A hardware/software framework for real-time spiking systems. In Wlodzislaw Duch, Janusz Kacprzyk, Erkki Oja, and et al., editors, *Artificial Neural Networks: Biological Inspirations — ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part I*, volume 3696 of *Lecture Notes in Computer Science*, pages 161–166. Springer-Verlag GmbH, Sep 2005.

N. Ouerhani, H. Hgli, P. Burgi, and P. Ruedi. A real time implementation of the saliency-based model of visual attention on a simd architecture. In Springer, editor, *Lecture notes in computer science. Pattern Recognition : 24th DAGM Symposium, Zurich, Switzerland, September 16-18, 2002. Proceedings*, volume 2449/2002, pages 282–289. Springer Berlin / Heidelberg, 2002.

S. Park, S. Ban, Shin J., L. Minho, O. Kaynak, E. Alpaydin, E. Oja, and X. Lei. Implementation of visual attention system using bottom-up saliency map model. In Berlin Springer, editor, *Lecture notes in computer science, Artificial neural networks and neural information processing - ICANN / ICONIP*, volume 2714, pages 678–685, June 2003.

D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.

D.J. Parkhurst and E. Niebur. Texture contrast attracts overt visual attention in natural scenes. *European journal of neuroscience*, 19:783–789, 2004. doi: 10.1111/j.1460-9568.2003.03183.x.

R. Perfetti. "winner-take-all" circuit for neurocomputing applications. In *IEE Proceedings*, volume 137, October 1990.

L. Pessoa, S. Kastner, and L.G. Ungerleider. Neurimaging studies of attention: From modulation of sensory processing to top–down control. *Journal of neuroscience*, 23 (10):3990–3998, May 2003.

M.I. Posner. Orienting of attention. *Q J Experimental Psychology*, 32(1):3–25, 1980.

M.I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D. Bouwhuis, editors, *Attention and Performance*, volume X, pages 531–556. Erlbaum, Hillsdale, 1984.

W. Rall. Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input. *Journal of neurophysiology*, 30(5):1138–1168, September 1967.

J. Ramirez-Angulo, M. Robinson, and Sanchez-Sinencio. Current–mode continuous-time filters: Two design aprroaches. *IEEE Transactions on Circuits and Systems*, 39(6):337–341, June 1992.

C. Rasche. *The Making Of A Neuromorphic Visual System*. Springer, 2005.

C. Rasche and R.J. Douglas. Silicon synaptic conductances. *Journal of Computational Neuroscience*, 7:33–39, 1999.

C. Rasche and R. Hahnloser. Silicon synaptic depression. *Biological Cybernetics*, 84(1):57–62, 2001.

C. Rasche, R. Douglas, and M. Mahowald. Characterization of a pyramidal silicon neuron. In Smith L. S. and Hamilton A., editors, *Neuromorphic Systems: Engineering Silicon from Neurobiology*. World Scientific, 1997.

J.H. Reynolds and L. Chelazzi. Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27:611–647, 2004. doi: 10.1146/annurev.neuro.26.041002.131039.

J.H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, 19(5):1736–1753, March 1999.

J.H. Reynolds, T. Pasternak, and R. Desimone. Attention increases sesitivity of v4 neurons. *Neuron*, 26:703–714, June 2000.

M. E. Robinson, H. Yoneda, and E. Sńchez-Sinencio. A modular CMOS design of a Hamming network. *IEEE Transactions on neural networks*, 3(2):444–456, May 1992.

J. Ross, M.C. Morrone, M.E. Goldberg, and D.C. Burr. Changes in visual perception at the time of saccades. *Trends in Neurosciences*, 24(2):113–121, February 2001. URL `http://www.sciencedirect.com/science/article/B6T0V-4599DYM-R/2/a0da41f508c58f73c9a17cef33381cfd`.

L.C. Rutherford, S.B. Nelson, and G.G. turrigano. Bdnf has opposite effects on the quantal amplitude of pyramidal neuron and interneuron excitatory synapses. *Neuron*, 21:521–530, September 1998.

E. Salinas and L.F. Abbott. A model of multiplicative neural responses in parietal cortex. *Proc. Natl. Acad. Sci.*, 93:11956–11961, October 1996.

R Sarpeshkar. Brain power – borrowing from biology makes for low power computing – bionic ear. *IEEE Spectrum*, 43(5):24–29, May 2006.

R. Sarpeshkar. Analog versus digital: Extrapolating from electronics to neurobiology. *Neural Computation*, 10(7):1601–1638, October 1998.

S Satyanarayana, Y.P. Tsividis, and H.P. Graf. A reconfigurable VLSI neural network. *IEEE Jour. Solid-State Circuits*, 27(1):67–81, Jan 1992.

S. R. Schultz and M. A. Jabri. Analogue VLSI 'integrate-and-fire' neuron with frequency adaptation. *Electronic Letters*, 31(16):1357–1358, Aug 1995.

E. Seevinck. Companding current-mode integrator: A new circuit principle for continuous-time monlithic filters. *Electronics Letters*, 26(24):2046–2047, November 1990.

R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gómez-Rodríguez, H. Kolle Riis, T. Delbrück, S. C. Liu, S. Zahnd,

A. M. Whatley, R. J. Douglas, P. Häfliger, G. Jimenez-Moreno, A. Civit, T. Serrano-Gotarredona, A. Acosta-Jiménez, and B. Linares-Barranco. AER building blocks for multi-layer multi-chip neuromorphic vision systems. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, Dec 2005.

T. Serrano-Gotarredona and B. Linares-Barranco. A high-precision current-mode WTA-MAX circuit with multichip capability. *IEEE journal of solid -state circuits*, 33(2):280–286, February 1998.

M. Shams, J. C. Ebergen, and M. I. Elmasry. Modeling and comparing CMOS implementations of the c–element. *IEEE Transactions on VLSI Systems*, 6(4):563–7, 1998.

R.Z. Shi and T.K. Horiuchi. A summating, exponentially-decaying CMOS synapse for spiking neural systems. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004a.

R.Z. Shi and T.K. Horiuchi. A VLSI model of the bat lateral superior olive for azimuthal echolocation. In *Proceedings of the 2004 International Symposium on Circuits and Systems (ISCAS04)*, volume 4, pages 900–903, May 2004b.

S. Sol'is-Bustos, J. Silva-Martínez, F. Maloberti, and E. Sánchez-Sinencio. A 60-db dynamic-range CMOS sixth-order 2.4-hz low-pass filter for medical applications. *IEEE Transactions on Circuits and Systems II*, 47(12):1391–1398, Dec. 2000.

M. Srinivasan, S. Zhang, M. Lehrer, and T. Collett. Honeybee navigation en route to the goal: visual flight control and odometry. *J Exp Biol*, 199(1):237–244, 1996. URL `http://jeb.biologists.org/cgi/content/abstract/199/1/237`.

A. Starzyk, J. and X. Fang. CMOS current mode winner-take-all circuit with both excitatory and inhibitory feedback. *Electronic Letters*, 29(10):908–910, May 1993.

A. Stocker and R. Douglas. Computation of smooth optical flow in a feedback connected analog network. In *Advances in Neural Information Processing Systems*, volume 12, 1998.

T.L. Taylor. Inhibition of return for expected and unexpected targets. *Acta Psychologica*, 2006. doi: doi:10.1016/j.actpsy.2006.03.006.

K.G. Thompson, N.P. Bichot, and T.R. Sato. Frontal eye field activity before visual search errors reveals the integration of bottom–up and top–down salience. *Journal of Neurophysiology*, 93:337–351, January 2005a. doi: 10.1152/jn.00330.2004.

K.G. Thompson, K.L. Biscoe, and R.T. Sato. Neuronal basis of covert spatial attention in the frontal eye field. *Journal of neuroscience*, 25(41):9479–9487, October 2005b. doi: 10.1523/jneurosci.0741-05.2005.

A. M. Treisman and G. Gelade. A feature–integration theory of attention. *Cognit Psychol*, 12:97–136, 1980.

M.V. Tsodyks and H. Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *PNAS*, 94:719–723, January 1997.

G.G. Turrigiano. Homeostatic plasticity in neural networks: the more things change, the more they stay the same. *Trends in Neuroscience*, 22(5):221–227, 1999.

G.G. Turrigiano, K.R. Leslie, N.S. Desai, L.C. Rutherford, and S.B. Nelson. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391: 892–896, February 1998.

A. van Schaik. Building blocks for electronic spiking neural networks. *Neural Networks*, 14(6–7):617–628, Jul–Sep 2001.

A. van Schaik and S.-C. Liu. AER EAR: A matched silicon cochlea pair with address event representation interface. In *IEEE International Symposium on Circuits and Systems*, volume V, pages 4213–4216, May 2005.

X. Wang. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *Journal of Neuroscience*, 19:9587–9603, November 1999.

K. Wawryn and B. Strzeszewski. Current mode AB class WTA circuit. In *The 8th IEEE International Conference on Electronics, Circuits and Systems, 2001. ICECS01*, volume 1, pages 293–296, 2001.

J.M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bullettin & review*, 1(2):202–238, 1994.

J.M. Wolfe and T.S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Review Neuroscience*, 5(4):495–501, June 2004.

K.A. Zaghloul and K. Boahen. A silicon retina that reproduces signals in the optic nerve. *Journal of Neural Engineering*, 3:257–267, December 2006. doi: 10.1088/1741-2560/3/4/002.