Feature Competition in a Spike-Based Winner-Take-All VLSI Network

Shih-Chii Liu and Matthias Oster
Institute of Neuroinformatics
University of Zürich and ETH Zurich
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
Email: shih,mao@ini.phys.ethz.ch

Abstract— **Recurrent networks and hardware analogs that perform a winner-take-all computation have been studied extensively. This computation is rarely demonstrated in a spiking network of neurons receiving input spike trains. In this work, we demonstrate this computation not only within an aVLSI network but also across networks of integrate-and-fire neurons in a feature competition task. The chip has four populations of neurons receiving input spike trains that represent the outputs of four feature maps. The connectivity within each population is configured so that all the neurons compete with one another. In addition, a second level of competition, which we call the feature competition, can be introduced between all populations (or feature maps). The two levels of competition are useful in a system that has to select both the locations of relevant features and the best feature map that is coded in an input stimulus. The selection process can be completed as fast as after two input spikes.**

I. Introduction

The construction of multi-chip VLSI systems with large-scale networks of spiking neurons and spike-based sensors is rapidly becoming a reality. These systems rely on the robust operation of a flexible infrastructure that allows event-based communication between multiple chips [1], [2], [3], [4]. With prototypical technology in hand to construct such systems, we are starting to explore the performance of event-based systems in various processing tasks.

One computationally powerful operation that can be expressed by spike-based recurrent networks is the winner-take-all function [5]. This computation is intrinsic to many models that describe attention and recognition processes in the cortex [6], [7] and is thought to be a basic building block of the cortical microcircuit [5]. It have also be demonstrated in hardware analogs [8], [9], [10].

In this work, we describe how we used the winner-take-all function to implement feature competition in a chip (*'Object'* chip) which has four populations of aVLSI integrate-and-fire neurons. This chip is part of a multi-chip multi-layered asynchronous spike-based vision system (*CAVIAR*) that classifies spatio-temporal trajectories in the scene [11]. The components of this system all communicate using an asynchronous event-based transmission protocol called the address-event representation (AER) protocol. This mechanism routes spike events between neurons and synapses that are labelled by unique addresses [1], [2]. This protocol permits neurons/pixels to be virtually connected on or across chips. The system consists of a transient retina, four convolution chips, the *'Object'* chip, and a spike-based learning chip in sequence.

The *CAVIAR* system can implement an abstract version of a hierarchical "object recognition" system similar to that proposed by [12]. In their model, the authors used a MAX function to obtain translation and size invariant responses from their feature detectors. In a similar spirit, the feature maps in *CAVIAR* are created by convolving retinotopic input with preprogrammed feature kernels in the convolution chips.

The *Object* chip implements a spike-based version of the MAX function on the output of each convolution chip, thus obtaining a translation invariant feature map. By selecting the best feature map out of the maps generated by the convolution of the retinotopic input with the same feature detector but at different scales, the *Object* chip can also achieve size invariant responses. Since the kernels of the convolution stage are freely programmable, a variety of different vision algorithms can be explored with this architecture.

In this paper, we show how the *Object* chip implements feature competition across the emulated outputs of 4 different spike-based convolution chips which have been preprogrammed with different feature kernels. Different features can also correspond to the same feature detector at different scales.

The goal of this work is to show how this module can implement feature competition with asynchronous spiking neurons, and how this process can be completed in as few as two input spikes thus making the selection process fast [13]. The inputs to the network do not have to be spatially static. In fact, the *CAVIAR* system is designed to work with moving stimuli especially since the visual input is first processed by a transient retina chip [14]. The winning neurons in the *Object* chip code the highest input activity to the population. This chip reduces the data flow rate to the classifier chip by preserving only information about the best feature map. In the *CAVIAR* system, it also extracts the depth information by programming the same feature detector at different scales in the different convolution chips.

II. *Object* Chip

This AER transceiver chip was fabricated in a 0.35μ CMOS technology and consists of four populations of 8 x 8 VLSI integrate-and-fire neurons with various types of synapses. The

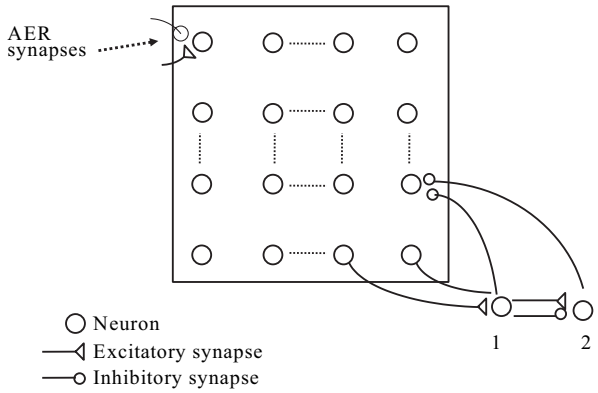

Fig. 1. Architecture of 'Object' chip configured for competition within a feature map and across feature maps. Each population of neuron consists of 62 excitatory neurons and two inhibitory neurons. Inhibitory neuron 1 is excited by all the excitatory neurons and in return, it inhibits these neurons. Inhibitory neuron 2 is used for competition across feature maps. It is excited and inhibited by inhibitory neuron 1 in its own population and excited by inhibitory neuron 1 of all other populations.

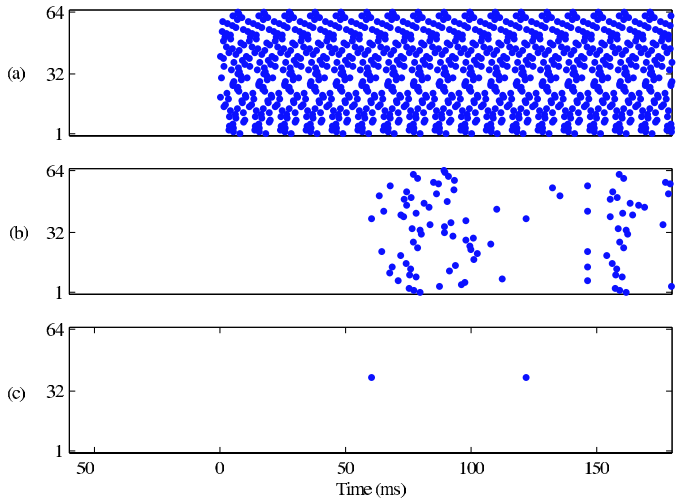

Fig. 2. Example raster plot of the spike trains to and from the neurons: (a) Input: starting from 0 ms, the neurons are stimulated with spike trains of a regular frequency of 100Hz, but randomized phase. Neuron number 42 receives an input spike train with an increased frequency of 120Hz. (b) Output without WTA connectivity: after an adjustable number of input spikes, the neurons start to fire with a regular output frequency. The output frequencies of the neurons are slightly different due to mismatch in the synaptic efficacies. Neuron 42 has the highest output frequency since it receives the strongest input. (c) Output with WTA connectivity: only neuron 42 with the strongest input fires, all other neurons are suppressed.

connectivity of the populations can be programmed through the AER infrastructure and the local connectivity on-chip.

The chip was designed so that it can receive inputs from up to 4 spike-based convolution chips with programmable feature kernels. The inputs from one convolution chip indicate the spatial locations of its preprogrammed feature kernel and the firing rates represent the strength of the convolution. The *Object* chip determines which feature map has the strongest outputs and in addition, computes the best spatial location of that feature. The processing on this chip reduces the image information rate to the subsequent post-processor.

The best feature map is determined by configuring the connectivity of each population so that it implements the *hard winner-take-all* function. This means that only one neuron in the population will be active as shown in Fig. 1. The winner and the global inhibitory neuron in each population code the input activity, that is, their output activities are proportional to the input activity.

A. Chip Architecture

The chip architecture in Fig. 1 shows one of the four populations of 8x8 integrate-and-fire neurons. The neuron circuit implements an integrate-and-fire model with a constant leak current. The externally controllable parameters for the neuron circuit include the threshold voltage, the refractory period, the pulse width of the spike, and the leak current. 62 out of the 64 neurons in each population are considered as excitatory neurons and the remaining two neurons are inhibitory neurons.

Each neuron has 8 AER input synapses of the excitatory, the excitatory depressing, and the inhibitory type. In addition, every neuron has 2 sets of local synapses. The first set consists of the connections that a neuron can make to other neurons and the second set consists of the connections that the neuron receives from other neurons. The type of connection made from one neuron to another is indicated by the type of neuron. The only exception is the connections made by global inhibitory neuron 1.

Each excitatory neuron connects to its four neighbors, and to inhibitory neuron 1 of its population. It receives excitatory connections from its neighbors and inhibitory connections from both global inhibitory neurons.

Inhibitory neuron 1 connects to all excitatory neurons. It also makes an on-chip excitatory and inhibitory connection to inhibitory neuron 2 and makes an excitatory connection to the inhibitory neuron 2 of the remaining populations. In return, it receives connections from all excitatory neurons of its population.

Inhibitory neuron 2 connects to all excitatory neurons of its population. In addition, it receives excitatory connections from inhibitory neuron 1 of all 4 populations and an inhibitory connection from inhibitory neuron 1 of its own population. The local synapses can be activated without going through the AER infrastructure.

The spiking activity of the neurons can be monitored through the addresses on the AER bus while an on-chip scanner allow us to monitor the membrane potentials of the neurons externally. Details of the neuron and synaptic circuits and the AER infrastructure supporting the programmability of the system have been previously described [3].

B. Connectivity Setup

The neurons express a certain amount of heterogeneity because of inherent nonidealities in the fabrication process. The variance in the response of the neurons can be corrected

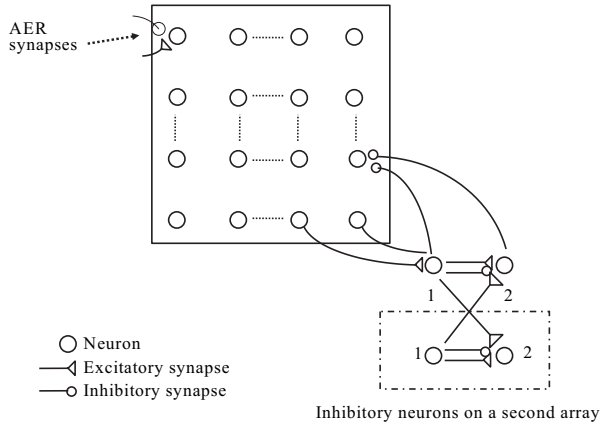

Fig. 3. Architecture of 'Object' chip configured for competition within two feature maps and competition across feature maps. All excitatory neurons of a winner-take-all network receive inputs from a feature map. These neurons excite a global inhibitory neuron which in turn inhibits all excitatory neurons. For competition across feature maps, inhibitory neuron 2 in each population is excited by inhibitory neurons 1 in all populations and it inhibits all excitatory neurons in its own population. Notice, that it is also inhibited by the output of inhibitory neuron 1 in its own population.

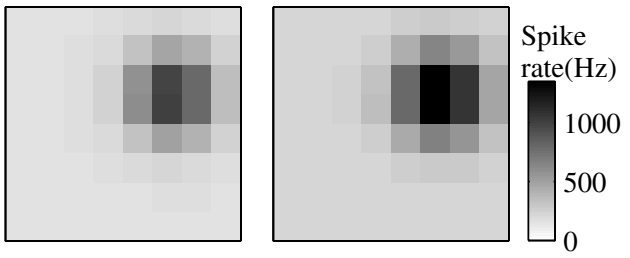

Fig. 4. Two of the four input feature maps for the experiment described in Section IV. Inputs to the neurons in each population consist of spike trains of regular frequency and represent a two-dimensional Gaussian distribution that rotates around the center of the array with a frequency of 0.25Hz. The distribution is scaled so that the highest input rate equals 1000Hz. This high rate is necessary because the network has to determine the winner using an estimate of the instantaneous input rates on a moving stimulus. In addition, every neuron receives a background firing of 200Hz so that all neurons spike at a low quiescent rate. The feature map on the right is presented to only one population while the one on the left is presented to the remaining 3 populations. The firing rates of the right feature map have been increased by 35%.

up to an average of 10% by using a spike burst encoding method [15] to reduce the mismatch in the synaptic weights across neurons.

To implement competition within the network, we activate the local excitatory connections from the population to its global inhibitory neuron 1 (see Fig. 1) and the inhibitory connection from this neuron to the population. The winner is selected after a pre-determined number of input spikes according to the constraints of the connectivity parameters needed for the WTA function [15] [16]. Each excitatory input spike charges the membrane of the post-synaptic neuron until one neuron in the array reaches threshold after the pre-determined number of input spikes and is reset. This neuron

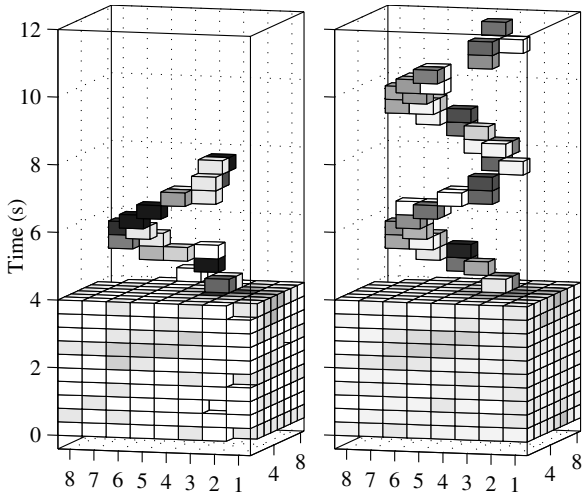

Fig. 5. Output of the 'Object' chip showing the competition across 4 feature maps. Only 2 feature maps are shown. Output of the chip is displayed over time for each neuron (column). Every box represents the spike rate encoded in the gray level. From $t=0$ to 4s, the internal connectivity is switched off and most of the neurons are active. At time $t=4$s, competition takes place within each feature map. All neurons except for the winner are suppressed. The position of the winning neuron follows the center of the Gaussian distributed input, which rotates once in 4s. At time $t=8$s, the feature maps compete with each other, leaving only the winning feature map with its most active (or winning) neuron (right side).

then drives the inhibitory neuron which in return inhibits all other neurons. Self-excitation of the winning neuron facilitates the selection of this neuron as the next winner.

To implement feature competition, we activate the on-chip connections from every inhibitory neuron 1 to all the four inhibitory neurons 2 of the four populations. In addition, we also activate the connections from each inhibitory neuron 2 to all excitatory neurons of its own population. Since the inhibitory neuron 1 codes the highest input activity within its own population, the one with the highest activity will suppress all other populations indirectly through their inhibitory neuron 2.

III. Competition Within Maps

We first demonstrate the WTA operation within one population of neurons. The network behaviour is illustrated using a spike raster plot in Fig. 2. At time $t=0$, the neurons receive inputs with the same regular firing frequency of 100Hz except for one neuron which receives a higher input frequency of 120Hz. The connectivity was configured in this one experiment so that the neuron reaches threshold in 6 input spikes, after which the network selects the neuron with the strongest input as the winner. On this chip, a network that has been compensated for synaptic mismatch can discriminate an average difference of input frequencies of 10% or an absolute minimum difference of around 20% across all neurons.

IV. Feature Competition

Next, we demonstrate feature competition by presenting each population with a rotating blob of Gaussian distributed

input on top of a background spike input rate of 200Hz as described in Fig. 4. This blob of activity is similar to the output of a convolution chip which sees an object rotating in the image. The background input is added so that all neurons have a low quiescent spike rate. For one of the populations, we increased its input spike rates by 35%. This was necessary because we did not compensate for the synapse mismatch in this experiment. As mentioned before, this percentage can be reduced to 20% if the synaptic mismatch is compensated. (We did not do the synaptic compensation here because the compensation scheme for all 4 populations would add a noticeable overhead to the communication rate on the AER bus.) The result of this competition across the populations is shown in Fig. 5. Initially, all four populations respond this rotating blob. At time $t = 4$s, the WTA competition in each population is initiated, and we see that only one neuron in each population stays active. The position of this neuron follows the center of the Gaussian distributed input. At time $t = 8$s, the feature competition across populations is initiated, resulting in only one active neuron in one active population.

V. Conclusion

We demonstrate feature competition in an aVLSI chip which has four arrays of integrate-and-fire neurons. Each array or population receives the outputs of a spike-based convolution chip which has a preprogrammed feature kernel. The output of a convolution chip represents a feature map of the input scene. By configuring the connectivity of each population for the hard winner-take-all operation, only the neuron receiving the highest input activity in each population will remain active, thus indicating the spatial location of the strongest output in that map.

The output activity of the global inhibitory neuron of each population also reflects the highest input activity to the population. Hence we are able to use the four inhibitory neurons from the four populations to compete against one another through an additional global inhibitory neuron in each population for the feature competition task. In this computation, only a single neuron in one population remains active while the other populations are suppressed.

This chip is part of a multi-chip multi-layered asynchronous spike-based vision system (*CAVIAR*) which classifies spatio-temporal trajectories in the image. It reduces the amount of information flow from the convolution chips to the classifier chip and also provides information continuously in time about the best feature map.

While the feature competition in our experiments could have been implemented in a single population that receives the outputs of all 4 feature maps, we intend to investigate ideas of, for example, normalization of the outputs of a single feature map before competition. The separation into the four populations allows us to pursue these questions.

The chip was fabricated in a 0.35μm CMOS 4-metal process and has an area of about 8.5mm^2. At present, we integrated the populations for four feature maps on a single chip but in a future implementation, each chip will be assigned to a single feature map. Competition across features will be implemented as competition across chips. This modification will allow the chips to be scaled easily according to the necessary number of feature maps [16] [11] [15].

VI. Acknowledgments

We acknowledge members of the INI hardware group for the development of the software and hardware infrastructure. This work is partially funded by EU-grant IST-2001-34124.

References

[1] S. R. Deiss, R. J. Douglas, and A. M. Whatley, "A pulse-coded communications infrastructure for neuromorphic systems," in *Pulsed Neural Networks*, W. Maass and C. M. Bishop, Eds. Boston, MA: MIT Press, 1999, ch. 6, pp. 157–178, iSBN 0-262-13350-4.

[2] K. A. Boahen, "Communicating neuronal ensembles between neuromorphic chips," in *Neuromorphic Systems Engineering*, T. Lande, Ed. Boston, MA: Kluwer Academic Publishers, 1998, ch. 11, pp. 229–259.

[3] S.-C. Liu and R. Douglas, "Spike synchronization in a network of silicon integrate-and-fire neurons," in *Proceedings of the 2004 IEEE International Symposium on Circuits and Systems*, vol. V, May 2004, pp. 397–400, ISCAS '04: Vancouver, Canada, 25 May–28 May.

[4] G. Indiveri, T. Horiuchi, E. Niebur, and R. Douglas, "A competitive network of spiking VLSI neurons," in *World Congress on Neuroinformatics*, F. Rattay, Ed. Vienna, Austria: ARGESIM/ASIM Verlag, Sept 24–29 2001, aRGESIM Reports.

[5] R. Douglas and K. Martin, "Cortical microcircuits," *Annual Review of Neuroscience*, vol. 27, no. 1f, pp. 419–451, 2004.

[6] C. Itti, E. Niebur, and C. Koch, "A model of saliency-based fast visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Apr 1998.

[7] D. Lee, C. Itti, C. Koch, and J. Braun, "Attention activates winner-take-all competition among visual filters," *Nature Neuroscience*, vol. 2, pp. 375–381, Apr 1999.

[8] J. Lazzaro, S. Ryckebusch, M. A. Mahowald, and C. A. Mead, "Winner-take-all networks of O(n) complexity," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1989, vol. 1, pp. 703–711.

[9] G. Indiveri, "Modeling selective attention using a neuromorphic analog VLSI device," *Neural Computation*, vol. 12, no. 12, pp. 2857–2880, 2000.

[10] R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, pp. 947–951, Apr 2000.

[11] R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, H. Riis, T. Delbrück, S.-C. Liu, P. Häfliger, G. Jimenez-Moreno, A. Civit, T. Serrano-Gotarredona, A. Acosta-Jimenez, and B. Linares-Barranco, "AER building blocks for multi-layer multi-chip neuromorphic vision systems," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006, vol. 18.

[12] M. Riesenhuber and T. Poggio, "Models of object recognition," *Nature Neuroscience*, vol. 3, pp. 1199–1204, 2000.

[13] D. Z. Jin and H. S. Seung, "Fast computation with spikes in a recurrent neural network," *Phys Rev E*, vol. 65, no. 5, pp. 051 922–1–051 922–4, 2002.

[14] P. Lichtsteiner and T. Delbrück, "64x64 event-driven logarithmic temporal derivative silicon retina," in *Proceedings of the 2005 IEEE Workshop on Charge-Coupled Devices and Advanced Imager Sensors*, June 2005, Nagao Prefecture, Japan, 9–11 June.

[15] M. Oster and S.-C. Liu, "A winner-take-all spiking network with spiking inputs," in *Proceedings of the 11th IEEE International Conference on Electronics, Circuits and Systems*, December 2004, ICECS '04: Tel Aviv, Israel, 13–15 December.

[16] ——, "Spiking inputs to a winner-take-all network," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006, vol. 18.