Spike-Based Vision Processing. Seeing without Frames

Where are we and where should we go?

Bernabé Linares-Barranco

Instituto de Microelectrónica de Sevilla (IMSE), Centro Nacional de Microelectrónica (CNM), Consejo Superior de Investigaciones Científicas (CSIC). Ed. CICA, Av. Reina Mercedes s/n, 41012 Sevilla, Spain.

I. INTRODUCTION. THE POWER OF SPIKE-BASED COMPUTING

Conventional vision sensing and processing is based on capturing sequences of still frames and process them (or a subset of them) one by one. For sophisticated vision processing usually a complicated sequence of 2D operations is required: different convolution operations, edge detections and extraction, computation of orientations, grouping of segments, contour detections and segmentations, morphological operations, object recognition. Each of these operations are computationally expensive since they are applied on all pixels of the captured and processed images.

The vision system of living animals does not operate on acquiring and processing sequences of still frames. Biological vision systems exploit spike-based processing. Also, biological brains are made with a slow 'technology' which uses neurons whose response times are in the order of milliseconds. Nonetheless, the computational power of biological brains overwhelmingly outperforms present day fastest computers for tasks like vision (or sensory processing in general). The reason for this could be, on one side the massive parallelism and high number of processing units, but on the other side could also be justified by a different type of visual information coding, transmission and computation, much more efficient than frame based processing.

In the already concluded european project CAVIAR[1] (IST-2001-34124, www.imse.cnm.es/caviar) we have performed preliminary explorations of the powerful principles of vision sensing and processing using spiking events, through the use of the incipient AER (Address Event Representation) technology.

Biological cortical structures are composed of a small number of processing layers (no more than 10), as shown in Fig. 1. Each neuron in a layer connects to a 'projection field' in the next layer. Each connection in the projection field has a given 'strength' or 'weight', and the set of weights of the projection field is fixed for a given layer (it does not change with position of sending neuron). Consequently, this is mathematically equivalent to convolution operations, where the kernel of the convolution is the set of weights of the projection field. In the topology of Fig. 1, electrical spikes are sent from neurons in a layer to the neurons in the next layer through their respective projections fields. Spikes are asynchronous, and each neuron in a layer decides when to send a spike, depending on the history of spikes it has received. Therefore, in such a cortical structure a wavefront of spikes, from the first layer (the sensing layer) to the last layer (the result layer), crosses the layers while performing complex operations, transformations, and recognition. And this can be very fast, even if the speed of one particular processing element (a neuron) is small.

As an illustration, consider the multi-layer system shown in Fig. 2. This is a very simplified version of a neocognitron [1] type of structure for character recognition. The neocognitron implements sequences of convolutions. In this particular case, it is designed to distinguish between characters 'A' and 'H', which can be of different sizes and shapes (slight deformations). The input consists of a sensing retina with spiking outputs (AER) of 16x16 pixels. This retina could be of the type reported in [2]. The pixels of the retina will fire either one single spike or none, depending on whether they are part of the stimulus character (either 'A' or 'H'). The boxes named 'ki', 'pi', 'U', and 'C' are 16x16 pixel AER convolution processors [3], which have been programmed with the kernels shown in Fig. 3. Kernel 'U' is not shown because it is the unitary kernel (size 1x1). Each pixel in a convolution processor is an integrate-and-fire neuron which integrates the incoming events and generates an output spike when the integral reaches a threshold. Boxes 'Sp' are splitters, which copy their input

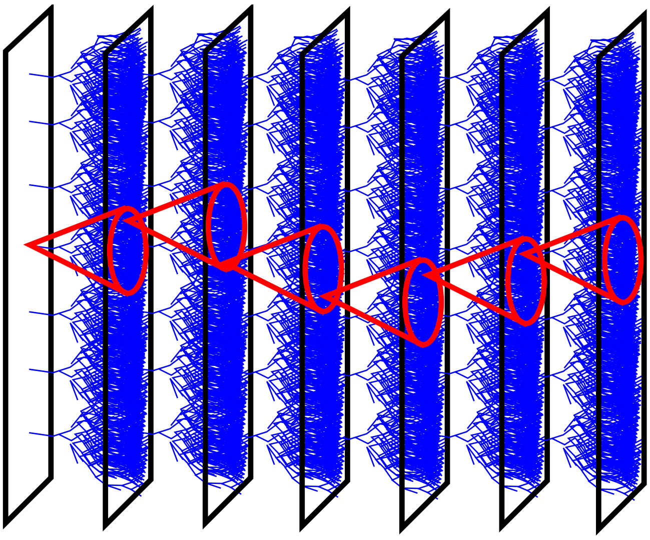


Fig. 1: Illustration of projection field concept in the brain. Each neuron in one layer connects to a projection field of neurons in the following layer. The weights of the connections follow a pattern which is independent of neuron position within a sending layer. Consequently, this is like applying a convolution from layer to layer.

[1] From June 2002 through November 2006.

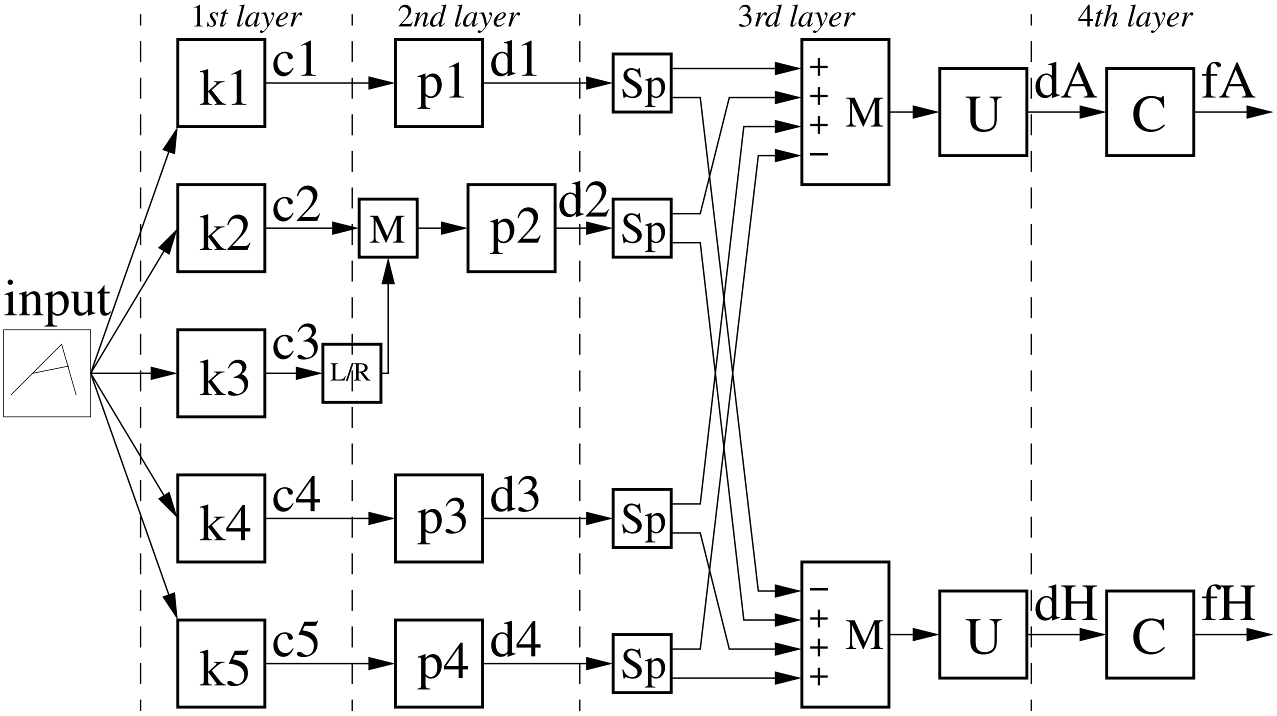


Fig. 2: Illustration of a multi-chip multi-layer AER convolution processing systems to distinguish between handwritten characters 'A' and 'H'. This system is loosely inspired in the neocognitron architecture.

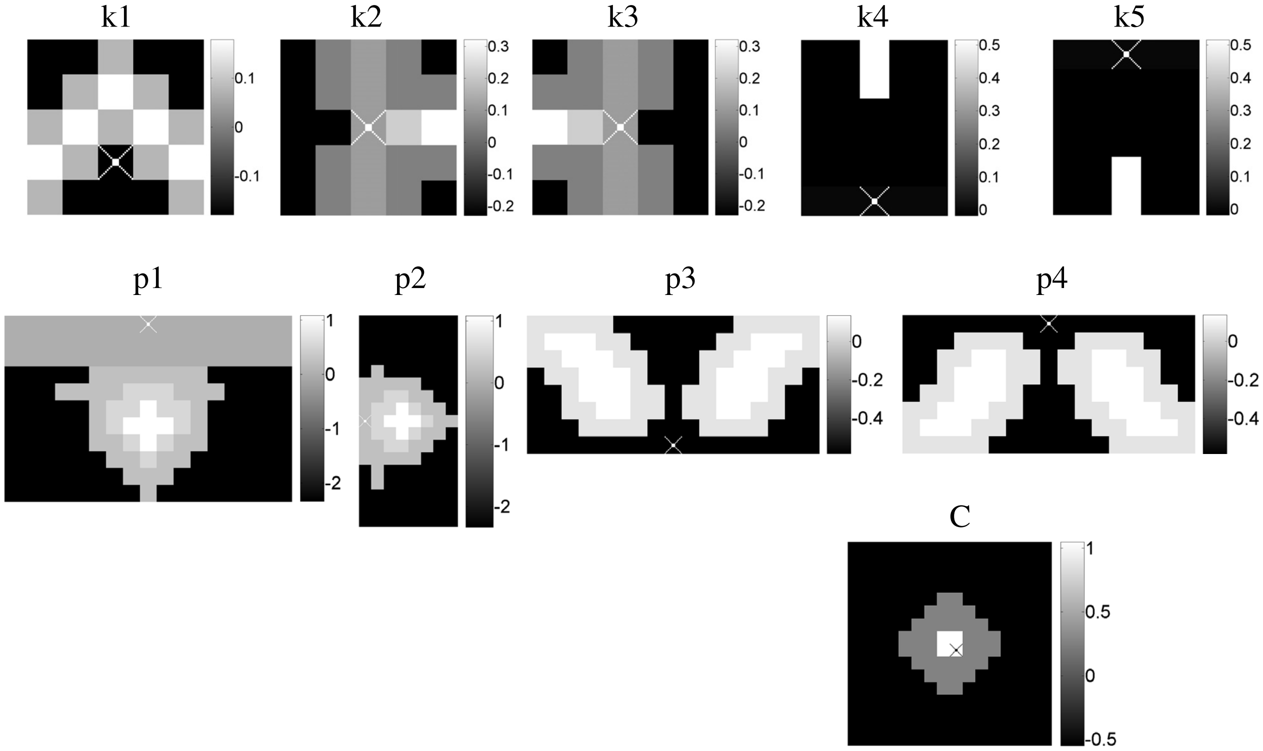


Fig. 3: Kernels used for the different convolutions in Fig. 2. The bar to the right side of each kernel shows the gray scale coding of the kernel value. For each kernel 'white' is assigned to its maximum value and 'black' to its minimum. Kernel values are normalized with respect to the threshold of the integrate-and-fire circuit. For example, if kernel value is '1' then one single positive event for this pixel would produce an output event. If kernel value is 0.5, then two positive events would be needed for an output event to be generated.

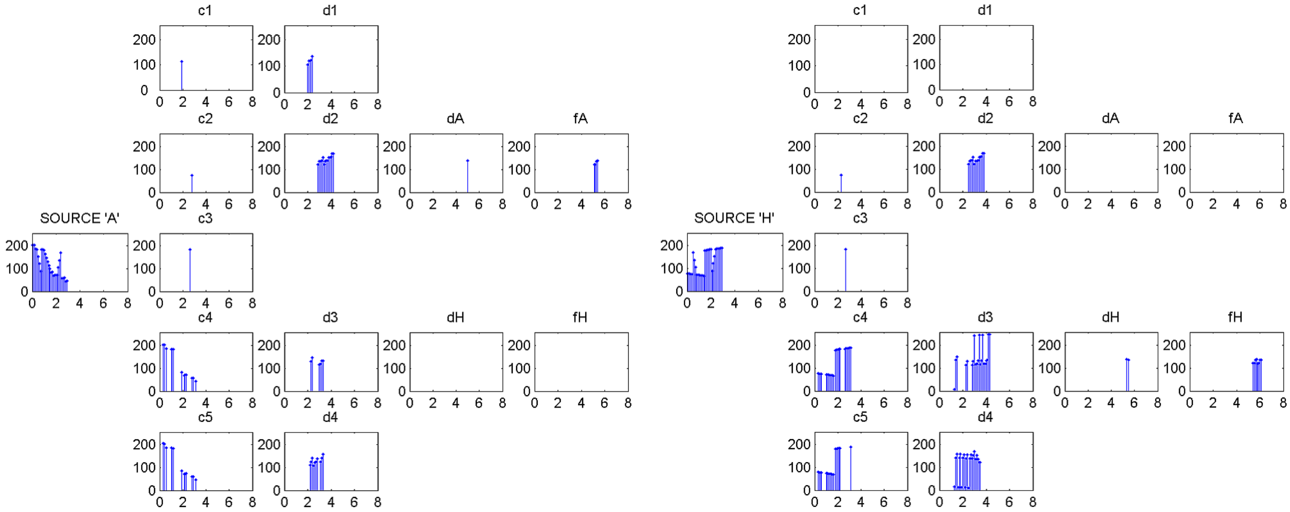


Fig. 4: Timing of the output events produced by the different convolution stages. Vertical axes represent pixel number in the 16x16 array (from 0 to 255), and horizontal axes are in μ*s*. The left side corresponds to the case of input stimulus ‘A’, while the right hand side corresponds to input stimulus ‘H’.

events on different output nodes. Boxes ‘M’ are mergers, which sequence input events coming from separate nodes onto a single output node. Other blocks are possible, like mappers, which perform transformations on the events. This is the case of the L/R block, which performs a left/right mirror operation on the events. The resulting flow of events on the different nodes can be seen in Fig. 4, for the cases of presenting either a character ‘A’ or ‘H’ at the input. Horizontal axes represent ‘time’ in micro-seconds. In this example, each block in Fig. 2 introduces a conservative delay of 100*ns*. By looking at the timing in Fig. 4, one can see that the delay between input stimulus presentation and correct output recognition is around 3μ*s*, although 13 convolutions have been performed, some of them with kernel sizes of up to 16x16. Such performance cannot be achieved with any present-day computing hardware, not even with very specialized digital convolution processors. For example, Öwall et al. [4] have reported a dedicated convolution processors which can do 15x15 kernels on images of size 256x256 in 55*ms*. Scaling this down to images of size 16x16 results in a delay of 215μ*s* for one single convolution.

Furthermore, when scaling up a structure like the one shown in Fig. 2, for example to handle a larger alphabet, what happens is that the number of blocks per layer grows, but not much the total number of layers. Consequently, the global computational delay will not grow significantly. For example, the handwritten character recognition system reported in [1] is composed of 8 layers with a total of 376 blocks, and is capable of recognizing 35 different handwritten alphanumeric characters. The expected computational delay for such a system would be of the order of twice that of Fig. 2, since the number of layers is twice, and the expected number of spikes per node should be similar.

II. CAVIAR DEVELOPMENTS ON AER SPIKING HARDWARE FOR VISION

In the EU funded research project CAVIAR (Convolution AER Vision Architecture for Real Time), four different european research groups joint during 4.5 years (June 2002 through November 2006) to develop chips and PCBs for demonstrating a simple multi-chip multi-stage spiking vision sensing/processing/actuating AER based system. The structure of the system is shown in Fig. 5. It consists of:

1.- A motion sensing retina (temporal contrast) with AER output [2].

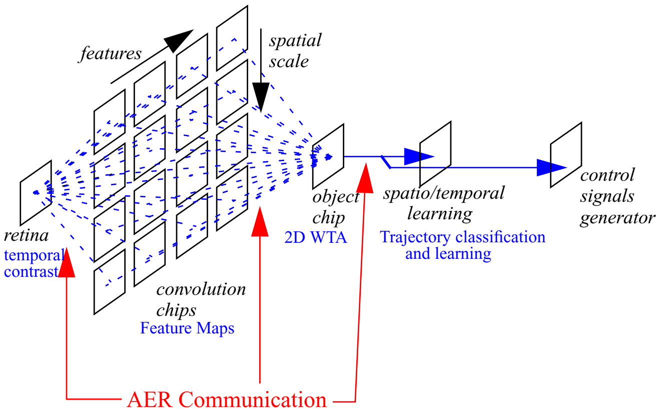


Fig. 5: AER Vision System developed in the CAVIAR project. It consists of a temporal contrast retina with AER output, followed by an array of AER convolution chips with programmable kernels of arbitrary shape and size, followed by a 2D AER Winner-Takes-All position and feature competition chip, followed by an AER trajectory learning/classification stage. In CAVIAR the system is configured to detect and follow balls. The control stage at the end provides control signals to move the retina towards the moving balls.

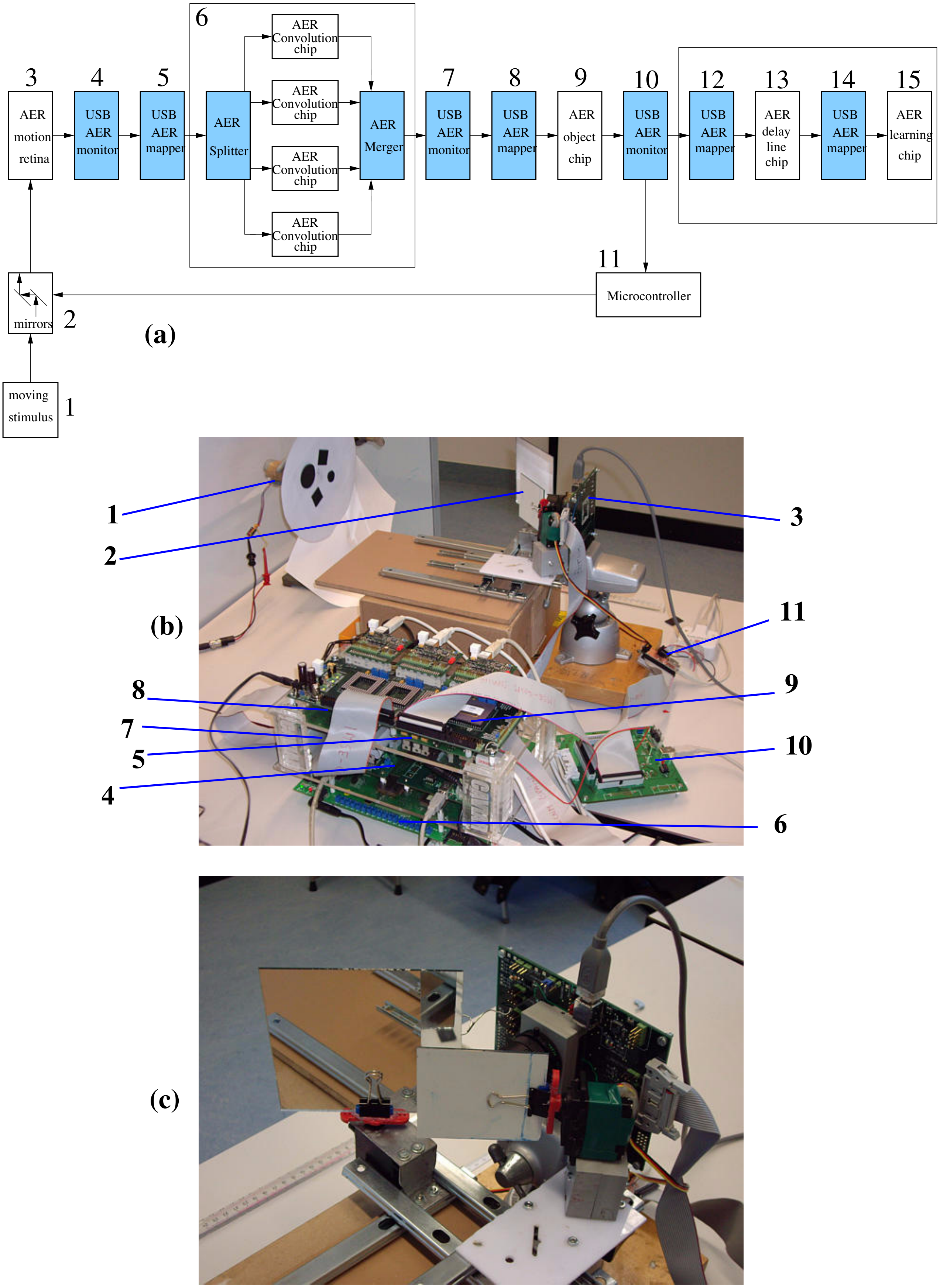
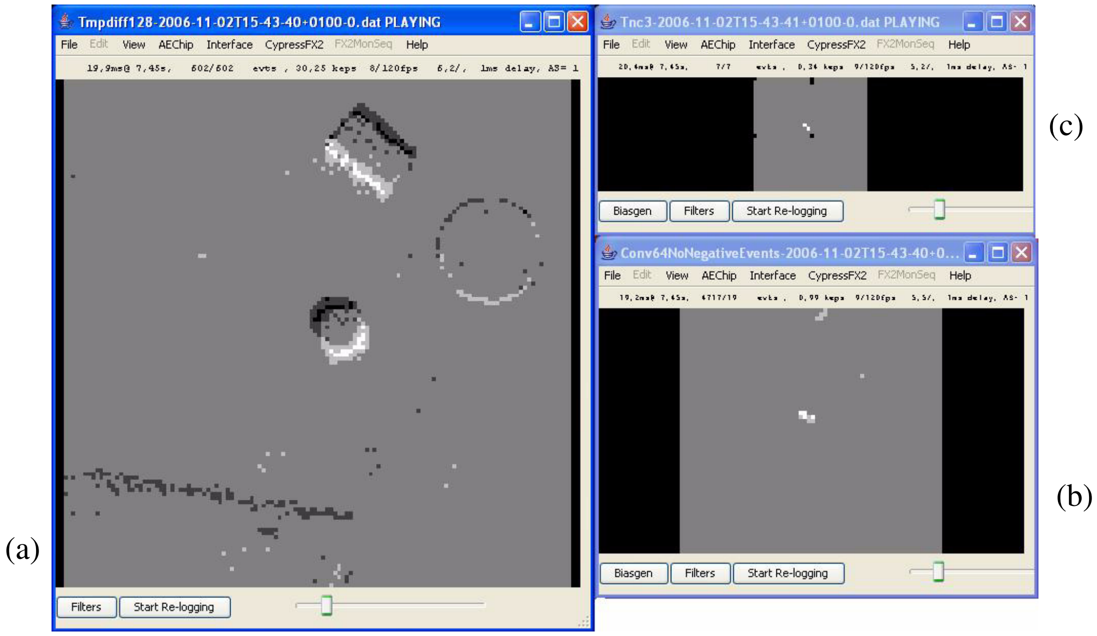


Fig. 6: Experimental setup of multi-layered AER vision system for ball tracking (white boxes include custom designed chips, blue boxes are interfacing PCBs). (a) block diagram, (b) photograph of setup, (c) details of mirrors.



Tmpdiff128-2006-11-02T15-43-40+0100-0.dat PLAYING

19,9ms@ 7,45s, 602/602 evts , 30,25 keps 8/120fps 6,2/, 1ms delay, AS= 1

Tnc3-2006-11-02T15-43-41+0100-0.dat PLAYING

20,4ms@ 7,45s, 7/7 evts , 0,34 keps 9/120fps 5,2/, 1ms delay, AS= 1

Conv64NoNegativeEvents-2006-11-02T15-43-40+0...

19,2ms@ 7,45s, 4717/19 evts , 0,99 keps 9/120fps 5,5/, 1ms delay, AS= 1

Fig. 7: AER flow monitored on different nodes of the CAVIAR Vision System. (a) 128x128 pixel output of temporal contrast retina. (b) 64x64 pixel output of convolution chips. (c) 32x32 pixel output of WTA chip.

2.- A set of AER convolution chips [3] with programmable kernels of arbitrary shape and size (up to 32x32). Convolution chips will be programmed to detect different features, and features of different scales. In the CAVIAR demonstrator, only one feature was used (a circular object) with up to four different sizes.

3.- A 2D AER Winner-Takes-All (WTA) chip [5] which for each convolution output (feature and scale) will detect its maximum, and from all maximums will select the absolute maximum.

4.- An AER trajectory learning/classification stage [6], which first produces delayed versions of the WTA outputs, which are then fed to an associative learning/ classification AER chip.

5.- Besides these chips, a rich set of AER chip-computer and chip-chip interfaces have been developed based on FPGAs [7]-[8], for (1) monitoring AER traffic of AER links and visualize them in real time on computer screens, (2) generating synthetic AER traffic from a computer and inject it as input to an AER stage, (3) remap events going from one chip to another, (4) log and replay events, and (5) adapt the AER signals for the (mechanical) control of the retina.

A preliminary version of the CAVIAR system was reported at the NIPS 2005 conference [9]. More information, including publications and videos, can be retrieved from the project web site at `http://www.imse.cnm.es/caviar`.[2]

The final CAVIAR demonstration system could discriminate and track balls of different sizes. A block diagram of the complete system is shown in Fig. 6(a), and a photograph of the complete experimental setup is given in Fig. 6(b). The complete chain consisted of different AER modules (chips and PCBs), all numbered in Fig. 6: (1) The rotating wheel stimulus. (2) Two moving mirrors for changing the visual direction of the retina. (3) The retina. The retina looked at a rotating disc with figures on it. (4) An AER monitor PCB which sends copies of the events to a computer for visualization purposes. (5) A USB-AER board as mapper to reassign addresses and eliminate the polarity of brightness change. (6) The convolution-PCB, which includes four 32x32 pixel convolution chips, an AER splitter, and an AER merger. The four chips can be programmed to operate as a single convolution processor of 64x64 pixel programmed with a unique kernel, or they can be programmed as four independent convolution processors of 32x32 pixels each with an independent kernel. (7) An AER monitor. (8) An AER mapper. (9) A PCB containing the 'object chip' which performs a winner-takes-all operation on the flow coming from the convolutions. (10) Another AER monitor. (11) A microcontroller progammed to control the mirrors in such a way that the figures detected by the visual chain will be centered in the visual field. (12) The output of the 'object chip' is transformed using a mapper (12) and fed to a delay line chip (13), the outputs of which are fed through a mapper (14) to a learning (15) chip.

The system retina looks at a rotating disk which has black figures drawn on it: two circles of different sizes and one or two rectangles. The retina senses moving edges. A monitor PCB collects the events, which can be seen (for a snapshot) in Fig. 7(a). The retina has 128x128 pixels. The convolution chips 'sees' all retina pixels but only provides output for the central 64x64 pixels. The kernel was programmed to follow the small size circle, and the output of the convolution stage can be seen in Fig. 7(b). This output is mapped from a 64x64 space to a 32x32 one, and processed by the WTA chip, which filters out noise, and gives clear strong events for the central coordinates of the small circle. This is shown in Fig. 7(c). The AER output

[2] The project full Final Report can be downloaded from 'Restricted Area2' after receiving username and password from the project coordinator.

of the WTA is used to act on two servo motors which act on two orthogonally controlled mirrors (see Fig. 6(c)) which change the angle of view of the retina. The system keeps the small circle centered on the visual field. Alternatively, another non-mechanical servo mechanism was developed, which changes the angle of view of the retina by introducing a controlled offset directly on the retina output events. These offsets are introduced by means of an AER mapper after the retina. This fully electronic means of actuating rendered much faster response capabilities, since now there were no mechanical parts involved (servos nor mirrors).

III. CONCLUSIONS AND FUTURE WORK

CAVIAR has demonstrated the viability and high potential of AER based multi-chip and multi-layer bio-inspired cortical-like spike-based (vision) sensing and processing systems. So far, only a few AER building blocks have been reported (by the CAVIAR consortium and other researchers world wide). There are AER sensing retinae for contrast and motion. CAVIAR has developed for the first time a fully programmable AER convolution chip without any restriction on kernel shape nor size. CAVIAR has also developed a 2D feature competition AER chip. CAVIAR has developed an associative learning/ classification AER chip. And CAVIAR has also developed a powerful set of AER chip-chip and chip-computer interfaces.

But CAVIAR has also managed to open a quite new way of conceiving very powerful, compact, sophisticated, and impressively fast (vision) sensing and processing hardware systems. CAVIAR has demonstrated that it is viable to work towards a computing paradigm based on what is known and what is yet to be discovered about biological brains and cortical structures. At this moment, future developments towards the consolidation of this bio-inspired computing paradigm, should address the following three aspects:

1) **A Theoretical Driven Aspect:** where research is performed towards developing a theory on how to assemble individual components, how to adjust their parameters and interconnections, to perform a specific desired functionality. Are new AER components required? How can global learning paradigms be introduced to train parameters? For the individual AER components, which parameters should be adjustable and which should be their ranges? In this respect, it would be interesting to develop a high-level behavioral simulator for such systems, which should be open enough to conceive new AER components, but specific enough to model non-idealities of the already available components. For this theory driven study, we should take into account what is already known in conventional image processing using convolutions: texture analyses, convolution based segmentations, wavelets for image processing, convolution neural networks (of which the neocognitron is a particular case and precursor), the work developed at Boston university on BCS (Boundary Contour System) which uses a sophisticated structure of convolutions for segmentation, and many others ... and try to adapt them to spike based processing.

2) **A Technological Driven Aspect.** It is clear that for building realistic cortical-like processing architectures, one will require a relative large number of individual components. The CAVIAR demonstrator, which can be considered a very simplistic single feature vision systems uses 8 AER chips plus 9 AER interfaces (see Fig. 6). The simplified neocognitron system in Fig. 2 uses 13 convolution processors plus 8 chip-chip interfaces, plus the sensor (monitors are not counted). A realistic neocognitron system, like the one described in [1], would use 8 cortical layers containing a total of 376 convolution blocks. In order to develop a reliable, robust, and feasible AER infrastructure that would allow researchers to ‘play’ with it and test realistic cortical structures, it should be conceived to be capable of hosting hundreds (or thousands) of individual components. It will be mandatory to miniaturize individual components to sizes of a few *cms*, which could be mounted and freely interconnected on some kind of mother-boards. Motherboards should be capable of hosting a high number of components (in the order of hundreds), they should be of compact size (like a laptop computer), and should be stackable. The infrastructure should be sufficiently open, so that as components are improved or newly conceived ones become available, compatibility is guaranteed.

3) **An Applications Driven Aspect.** As the theory develops and the technology becomes more efficient and powerful, the possible range of applications in real world problems will grow and grow as well. However, we believe it is crucial to start focusing on specific possible applications for which this spiking computing paradigm can already be applied. Working on specific applications and problems is an efficient catalyst which helps in developing the previous two aspects of theoretical and technological developments. At this point in time, it will be very beneficial to identify one, two or three applications which would drive the whole research. Applications should be selected based on difficult problems to solve with present state-of-the-art computing paradigms and hardwares. They should combine both, high speed requirements and heavy computational loads. Also, applications demanding compactness and low power solutions, while requiring sophisticated (vision) processing, could be interesting to consider.

References

[1] K. Fukushima and N. Wake, “Handwritten Alphanumeric Character Recognition by the Neocognitron,” *IEEE Trans. Neural Networks,* vol. 2, No. 3, pp. 355-365, May 1991.

[2] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128x128 120dB 30mW Asynchronous Vision Sensor that Responds to Relative Intensity Change," *2006 IEEE ISSCC Digest of Technical Papers*, pp. 508-509, San Francisco, 2006.

[3] R. Serrano-Gotarredona, T. Serrano-Gotarredona, A. Acosta-Jiménez, and B. Linares-Barranco, “A Neuromorphic Cortical-Layer Microchip for Spike-Based Event Processing Vision Systems,” *IEEE Trans. on Circuits and Systems-I: Regular Papers*, vol. 53, No. 12, pp. 2548-2566, December 2006.

[4] V. Öwall, M. Torkelson, and P. Egelberg, “A Custom Image Convolution DSP with a Sustained Calculation Capacity of >1GMAC/s and Low I/O Bandwidth,” *Journal of VLSI Signal Processing*, vol. 23, pp. 355-349, 1999.

[5] S-C. Liu and M. Oster, “Feature Competition in a Spike Based Winner-Take-All VLSI Network,” *Proc. of the 2006 IEEE Int. Symp. Circ. and Syst.*, (ISCAS’06), pp. 3634-363637, May 2006.

[6] P. Häfliger, “Adaptive WTA with an analog VLSI neuromorphic learning chip,” *IEEE Trans. on Neural Networks*, to be published.

[7] F. Gomez-Rodriguez, R. Paz-Vicente, et al, “AER Tools for Communications and Debugging,” *Proc. of the 2006 IEEE Int. Symp. Circ. and Syst.*, (ISCAS’06), pp. 3253-3256, May 2006.

[8] R. Paz-Vicente, A. Linares-Barranco, et al “PCI-Aer Interface for Neuro-Inspired Spiking Systems,” *Proc. of the 2006 IEEE Int. Symp. Circ. and Syst.*, (ISCAS 2006), pp. 3161-3164, May 2006.

[9] 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. Douglas, P. Häfliger, G. Jimenez-Moreno, A. Civit, T. Serrano-Gotarredona, A. Acosta-Jiménez, B. Linares-Barranco, “AER Building Blocks for Multi-Layers Multi-Chips Neuromorphic Vision Systems” *Advances in Neural Information Processing Systems*, vol. 18, Y. Weiss and B. Schölkopf and J. Platt (Eds.), (NIPS’06), MIT Press, Cambridge, MA, pp. 1217--1224, 2006 [http://books.nips.cc/papers/files/nips18/NIPS2005_0268.pdf].