Event-Driven Sensing and Processing for High-Speed Robotic Vision

L. A. Camuñas-Mesa, T. Serrano-Gotarredona, and B. Linares-Barranco
Instituto de Microelectrónica de Sevilla (IMSE-CNM),
CSIC y Univ. de Sevilla, Sevilla, SPAIN
Email: camunas@imse-cnm.csic.es

***Abstract*— We present here an overview of a new vision paradigm where sensors and processors use visual information not represented by sequences of frames. Event-driven vision is inherently frame-free, as happens in biological systems. We use an event-driven sensor chip (called Dynamic Vision Sensor or DVS) together with event-driven convolution module arrays implemented on high-end FPGAs. Experimental results demonstrate the application of this paradigm to implement Gabor filters and 3D stereo reconstruction systems. This architecture can be applied to real systems which need efficient and high-speed visual perception, like vehicle automatic driving, robotic applications in non-structured environments, or intelligent surveillance in security systems.**

Keywords—Address Event Representation (AER), Event-driven vision, Event-driven processing, Event-driven convolutions, Gabor filters, High-speed vision, Bio-inspired vision

I. INTRODUCTION

State of the art in artificial vision is based on video streams, by capturing sequences of images at a given "frame rate" and processing them frame after frame by computational algorithms. Frame-by-frame processing is CPU-hungry, which means that feedback mechanisms for real-time compact-volume and low-power applications are out of the question.

On the other hand, biological vision is frame-free: neither the eyes nor the brain use video frames. In biology, retina cells (pixels) send electric spikes (events) asynchronously to the brain through the optical nerve. Brain cells process these spikes through complex hierarchical structures to achieve, for example, shape size and pose invariant object recognition. There are no frames, but a continuous flow of spikes/events from the retina through the brain cortex. There is neither a clock to signal the start or the end of a frame. Each neuron decides autonomously when to send out a spike depending on the spatio-temporal collection of spikes/events received. This asynchronous frame-free sensing and processing is also called here "event-driven" (as opposed to "frame-driven"). It has been shown that very fast object recognition (in about 150ms) is performed by the human ventral stream visual system [1].

This means that information propagates as spikes in such a way that the neurons that spike have time to send just one spike. Consequently, this reveals a highly efficient signal encoding in the brain. Based on these findings, researchers world-wide have developed during the past decade a collection of event-driven sensor [2]-[7] and processor [8]-[11] chips. For example, the "Dynamic Vision Sensor" (DVS) [4]-[7] has become very popular among the neuromorphic engineering community and is being used in a variety of

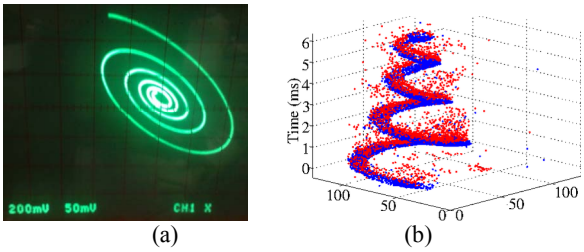

Fig. 1. Illustration of 128x128 pixel DVS operation. (a) 500Hz spiral on an analog oscilloscope in x-y mode observed by a DVS camera chip. (b) Recorded events from the DVS displayed in (*x,y,t*) coordinates. Color indicates sign of events: blue dots represent positive events, red dots represent negative events.

applications. Fig. 1 illustrates its basic principle, which helps to understand event-driven signal encoding. Each sensor pixel has an (*x,y*) coordinate or address. Whenever a pixel senses a change of light above a threshold it sends out of the chip a digital word (*x,y,s*) representing its address and a sign bit (positive for a dark to bright change and negative for a bright to dark change). This is called an event (or spike), and needs typically less than one micro second to be communicated. The DVS output consists of a flow of events. This is generally called AER "Address Event Representation". Fig. 1(b) shows the output events captured by one of our DVS chips during 6ms when observing the 500Hz spiral on an analog oscilloscope, shown in Fig. 1(a).

II. EVENT-DRIVEN PROCESSING AND PSEUDO-SIMULTANEITY PROPERTY

The practical implementation and usage of event-driven sensing and processing has revealed a very interesting property of these event-driven systems [10],[12] which is not possible for traditional frame-based vision, and which can completely revolutionize the artificial vision field and propagate to other fields. When performing event-driven sensing and processing, the input and output event flows of a processing stage are (in practice) simultaneous or coincident in time. We call this the **pseudo-simultaneity** property of event-driven processing. Event-driven processor chips process events as they flow in, with a delay per event typically in the 100ns range [10]. There is no need to wait for collecting image frames. Event-driven processors generate output events as they process the input event flow, as is done in brains. For example, for orientation extraction, as soon as enough events are received representing an oriented edge (typically 4 to 6 events are enough), a corresponding output event is produced, signaling the presence of an oriented edge in that location at that time.

This work has been funded by ERANET grant PRI-PIMCHI-2011-0768 (PNEUMA) funded by the Spanish Ministerio de Economía y Competitividad, Spanish research grants (with support from the European Regional Development Fund) TEC2009-10639-C04-01 (VULCANO) and TEC2012-37868-C04-01 (BIOSENSE), and Andalusian research grant TIC-6091 (NANONEURO).

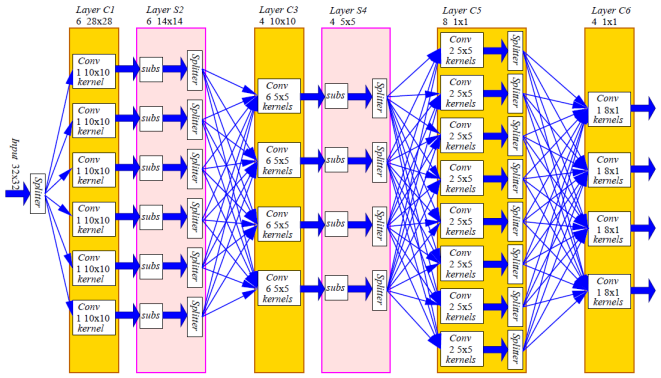

Fig. 2. Schematic block diagram of a multi-layer ConvNet

Thus, an output event flow (representing the 2D Gabor filter operation) is produced simultaneous to the input event flow (with the delay of a few events). This **pseudo-simultaneity** property holds also for multi-layer feed-forward processing. Fig. 2 shows schematically a camera sensor (retina) feeding a multi-layer structure of a feed-forward Convolutional Neural Network (ConvNet) [13], which is typically used for size and pose invariant object recognition, handwritten character recognition, scene recognition for robots, etc. If this ConvNet is implemented using traditional frame-driven sensing and image processing computing hardware [12], each stage has to wait until the output images from the previous stage are available. Fig. 3(b) shows the latencies in a frame-driven system when a symbol is quickly flashed (in 1ms) to the video camera sensor. A total of 6 frame delays are needed.

On the contrary, Fig. 3(c) shows the situation for an event-driven implementation. Thanks to the **pseudo-simultaneity** property, the output events of all stages are available concurrently to the sensor output event flow, and correct object recognition is feasible while the sensor is still producing events. This **pseudo-simultaneity** property has already been verified experimentally with cascades of available event-driven convolution chips [10], with large arrays of event-driven convolution modules implemented within high-end FPGAs [14], and has been verified by simulations of full feed-forward ConvNets processing high-speed DVS recordings, achieving symbol recognition with 1 to 2ms delays [15].

This example corresponds to the implementation of the network shown in Fig. 2 on an event-driven AER simulator to illustrate the high-speed capabilities of this sensing and processing approach. Fig. 4(a) shows an individual browsing a poker card deck, which can be fully browsed in less than one second. When recording such a scene with a DVS and playing the event flow back with jAER [18] one can freely adjust the frame reconstruction time and frame play back speed to observe the scene at very low speed. Fig. 4(b) illustrates a reconstructed frame when setting frame time to 5ms.

To provide a proper 32x32 pixel input scene to the network, we used an event-driven clustering-tracking algorithm [19] to track card symbols from the original 128x128 DVS recordings, since the instant they appeared until

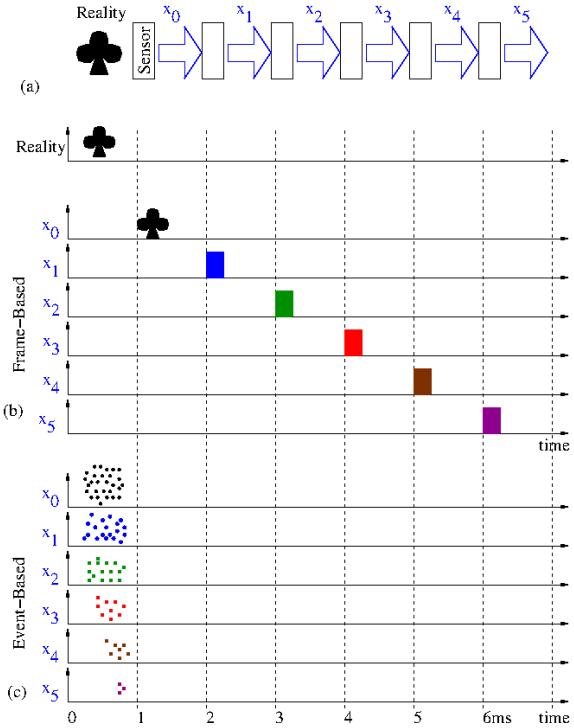

Fig. 3. Frame-driven versus event-driven feed-forward visual processing

they disappeared. Such time interval ranged typically from about 10-30ms per symbol and the 32x32 crop could contain on the order of 3k to 6k events. We sequenced several of these tracking crops containing all symbols and used the sequences as inputs to the Event-driven ConvNet in Fig. 2. We then tested the Event-driven ConvNet with the Event-driven AER simulator, obtaining a recognition success rate above 90% [15]. Fig. 4(c) shows an example situation of the output recognition events of the event-driven ConvNet while a sequence of 20 symbol sweeps was presented. The continuous line indicates which symbol was being swept at the input, and the output events are represented by different markers for each category: circles for "club" symbol, crosses for "diamond" symbol, inverted triangles for "heart" symbols, and non-inverted triangles for "spade" symbols.

III. Convolution Event-Driven Processing on FPGAs

Fig. 2 illustrates a multi-layer feed-forward Convolutional Neural Network (ConvNet). It contains a large number of convolution modules. Every time a pixel in a convolution module generates an output event, it is projected on a "projection field" in one or more modules (also called "feature maps") in the next layer. Event-driven convolutions are described in [8]-[10]: every time an event (x,y,s) is received by a convolution module, an event is sent to all pixels in that coordinate neighborhood which are updated in such a way that the event contribution is weighted spatially depending on the programmed convolution kernel. Depending on the module (or "feature map") of the previous layer, from which the event is coming, a different convolution kernel will be selected for the projection field [10]-[15].

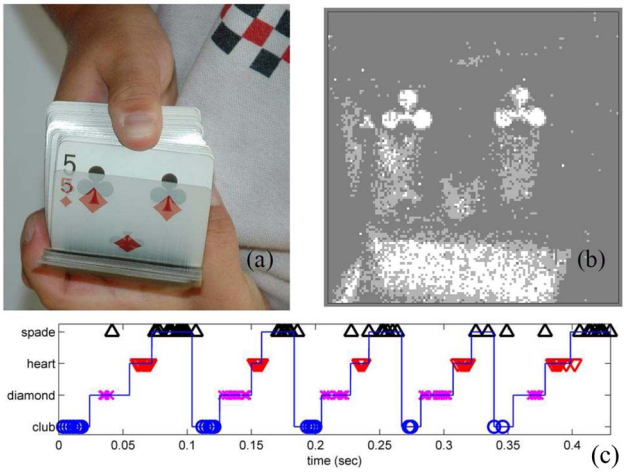

Fig. 4. Fast browsing of a poker card deck. (a) Picture taken with frame-driven camera. (b) jAER image obtained by collecting events during 5ms. (c) Recognition performance of the Poker Card Symbol Recognition ConvNet.

In order to map a 3D neural network multi-module topology, as in Fig. 2, onto hardware, we rely on network-on-chip (or network-on-board) concepts, as illustrated in Fig. 5. Each module in Fig. 2 is mapped onto an AER module in a 2D grid (Fig. 5(b)). Each module contains an event router, a configuration processor, and an event processor (like convolutions) (Fig. 5(c)). Each module in the grid is defined by a pair of indexes (*i,j*). The routing tables in each AER module define the connectivity of the network. Events travelling between modules contain the original event info (*x,y,s*) from the originating module, augmented by the coordinate of the module within the 2D grid. This way, a travelling event is defined as (*i,j,x,y,s*). Indexes (*i,j*) can be either the event source or destination module [14].

IV. Experimental Results

The event-driven sensing and processing approach has been tested using DVS sensor chips [7] together with 2D grid arrays of event-driven convolution modules implemented within high-end FPGAs. We have used it to perform 3D reconstruction in event-driven stereo vision.

A. Hardware setup

Fig. 5(a) shows the setup that we have used. The figure shows two DVS retina PCBs, each holding one DVS chip [7]. Each PCB sends address events through a parallel bus wire. The two DVS buses are connected to an event merger board [17]. This board arbitrates events of up to four inputs and merges all the event flow in one single address event bus, by adding extra bits to identify the source. This merger board is used in our setup whenever we want to use both DVS cameras simultaneously, as in stereo vision experiments. If we just need to use one DVS camera, the event merger board is bypassed. The merged event flow is sent to a custom-made convolutional board, where a 2D grid array of convolution modules is implemented within a Spartan6 FPGA [14]. Finally, a USBAERmini2 board [17] is used to timestamp all the events coming out of the convolutional board and send them to a computer though a high-speed USB2.0 port.

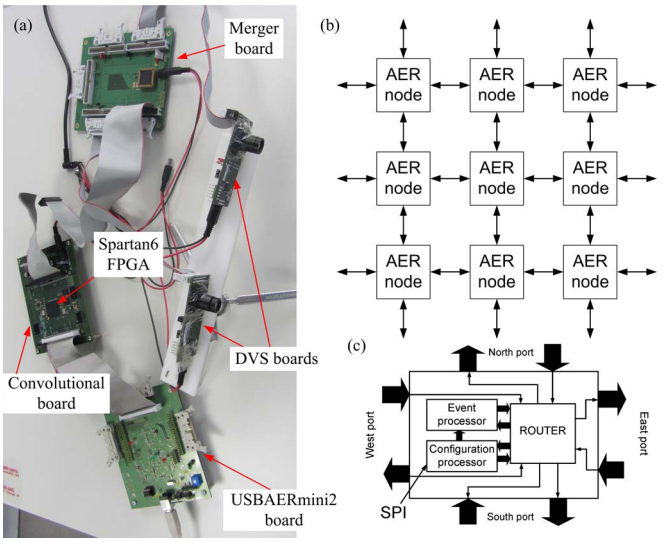

Fig. 5. (a) Example setup of a multi-sensor event-driven processing system using a Spartan6 FPGA and a family of custom-made event-handling boards. (b) 2D network topology for multi-casting-mesh AER. (c) Diagram of AER node.

B. 3D Stereo Reconstruction

First of all, we illustrate the operation of this setup with two simple experiments. In both we use visual input from one or two DVS sensors when observing a moving chart containing strong vertical and horizontal patterns, as well as thin 45º lines.

In a first experiment, only one DVS camera is used and the FPGA is programmed to hold an array of 3x3 convolution filters. The DVS events are fed to all nine convolution filters in the array, and each filter computes the corresponding convolution in parallel and event by event. The output events produced by each filter are then sent out through the corresponding routers to an output port of the FPGA board on a single parallel bus. Fig. 6(a) shows the simultaneous output of all nine Gabor filters when the retina is observing the moving stimulus. The kernels correspond to three different scales (kernel sizes) and three different angles. Fig. 6(a) has been generated using 80,000 events captured during 134ms.

In the second experiment both DVS sensors are connected to the merger board. This board adds an extra bit to each event to identify which retina produced the event. This way, the routers within the FPGA filter board can identify from which retina the event is coming from. If the event is coming from the left retina, the event is sent to the top three convolution modules and the central left one. If the event is coming from the right retina, it is sent to the three bottom modules and the central right one. The output from the convolution array within the FPGA is illustrated in Fig. 6(b). This figure has been generated by using 80,000 events, captured during 88ms.

In event-driven stereo vision, one key aspect is to identify the "corresponding events" between the two sensors. This is, for each event generated in one of the sensors (produced by some moving edge in reality), identify which is the corresponding event in the other sensor (produced by the same moving edge). In event-driven vision, we have the advantage that events are transmitted as they are produced.

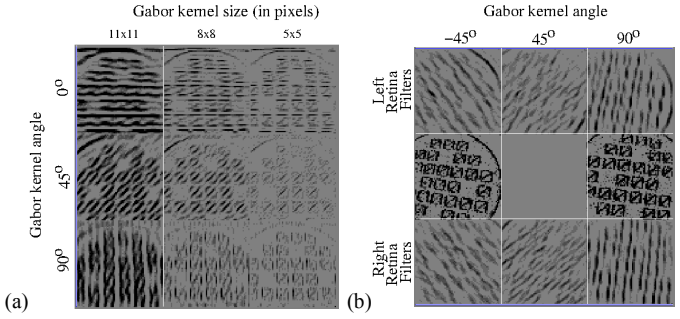

Fig. 6. Output of Event-Driven Gabor Filter Array. (a) One single DVS is used and events are sent to nine Gabor filters, processing them in parallel. (b) Stereo Vision setup: each retina output is sent to 3 Gabor filters of different angle plus another filter with unity kernel to replicate the input.

Consequently, the events produced by the same moving edge should happen at the same time, and the temporal correlation between events can be used as a clue to match events [20]. However, in practice there is some unavoidable jitter in event timing, and simultaneously produced events may be shifted some random time within some milliseconds range window [20]. Therefore, the event ordering differs between sensors. Geometric constraints are used to help identifying the matching events when a burst of events is received from both retinas within the same time window. In the present work, we use orientation of edges as a further clue to identify corresponding events [21]-[23]. Therefore, by implementing a bank of Gabor filters with different angles withing the Spartan6, we obtain a flow of oriented events which provides more reliably matched pairs of "corresponding events", improving the 3D reconstruction. Fig. 7 illustrates the results obtained after moving a ring in front of the DVS stereo setup.

CONCLUSIONS

A bio-inspired vision system with pseudo-simultaneous sensing and processing has been built. The system operation is based on a new event-driven computation paradigm. The system operation has been demonstrated first by implementing a multi-layer ConvNet in an event-based AER simulator, and finally in a binocular system able to extract edges simultaneously with the visual input to implement a 3D reconstruction algorithm. The high-speed response capabilities presented in these experiments confirm the applicability of this approach to robotic vision.

REFERENCES

[1] S. Thorpe, et al. "Speed of processing in the human visual system," Nature 381, 520-522, 1996.

[2] E. Culurciello, et al. "A biomorphic digital image sensor," IEEE J. Solid-State Circuits, 38:281-294, 2003.

[3] J. Costas-Santos, et al. "A spatial contrast retina with on-chip calibration for neuromorphic spike-based AER vision systems," IEEE Trans. Circuits Syst. I, 54,7:1444-58, 2007.

[4] P. Lichtsteiner, et al. "A 128x128 120 dB 15us latency asynchronous temporal contrast vision sensor," IEEE J. Solid-State Circuits, 43,2:566-576, 2008.

[5] C. Posch, et al. "A QVGA 143 dB Dynamic Range Frame-Free PWM Image Sensor With Lossless Pixel-Level Video Compression and Time-Domain CDS," IEEE J. Solid-State Circuits, 46,1: 259-275, 2011.

[6] J. A. Leñero-Bardallo, et al. "A 3.6 m latency asynchronous frame-free event-based dynamic vision sensor," IEEE J. Solid-State Circuits, 46,6:1443-1455, 2011.

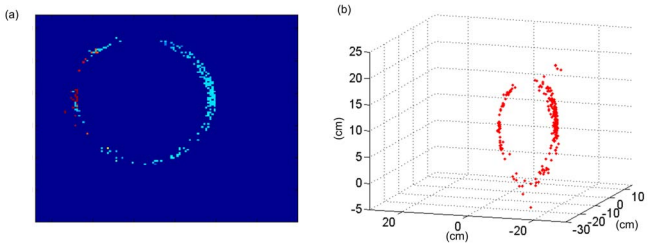

Fig. 7. (a) Disparity map reconstructed with a frame time of 50ms. (b) Result of the 3D reconstruction of the same frame.

[7] T. Serrano-Gotarredona and B. Linares-Barranco, "A 128x128 1.5% Contrast Sensitivity 0.9% FPN 3us Latency 4mW Asynchronous Frame-Free Dynamic Vision Sensor Using Transimpedance Amplifiers," IEEE J. Solid-State Circuits, vol.48, No. 3, pp. 827-838, March 2013.

[8] R. Serrano-Gotarredona, et al., "A Neuromorphic Cortical-Layer Microchip for Spike-Based Event Processing Vision Systems," IEEE Trans. Circuits and Systems, Part-I: Regular Papers, vol. 53, No. 12, pp. 2548-2566, December 2006.

[9] L. Camuñas-Mesa, et al. "A 32x32 Pixel Convolution Processor Chip for Address Event Vision Sensors with 155ns Event Latency and 20Meps Throughput," IEEE Trans. Circ. and Syst. Part-I, vol. 58, No. 4, pp. 777-790, April 2011.

[10] L. Camuñas-Mesa, et al. "An Event-Driven Multi-Kernel Convolution Processor Module for Event-Driven Vision Sensors," IEEE J. Solid-State Circ. 47,2:504-517, 2012.

[11] T.Choi, et al. "Neurom. Impl. of orientation hypercolumns," IEEE Trans. Circ. Syst. I,52,6:1049-60, 2005.

[12] C. Farabet, et al. "Large-scale FPGA-based convolutional networks," in Machine Learning on Very Large Data Sets, R. Bekkerman, M. Bilenko, and J. Langford (Eds) Cambridge Univ. Press, 2011.

[13] Y. LeCun, et al "Backprop. applied to handwritten zip code recogn.," Neur. Comp., 1,4:541-551, 1989.

[14] C. Zamarreño-Ramos et al. "Multi-Casting Mesh AER: A Scalable Assembly Approach for Reconfigurable Neuromorphic Structured AER Systems. Application to ConvNets," IEEE Trans. on Biom. Circ. Syst., vol. 7, No. 1, pp. 82-102, Feb. 2013.

[15] J. A. Pérez-Carrasco, et al. "Mapping from Frame-Driven to Frame-Free Event-Driven Vision Systems by Low-Rate Rate-Coding and Coincidence Processing. Application to Feed-Forward ConvNets," IEEE Trans. Patt. Anal. Mach. Intelligence, vol. 35, No. 11, pp. 2706-2719, Nov. 2013.

[16] T. Serrano-Gotarredona, et al "Very Wide Range Tunable CMOS/Bipolar Current Mirrors with Voltage Clamped Input," IEEE Trans. Circuits and Systems (Part I): Fundamental Theory and Applications, pp. 1398-1407, November 1999.

[17] R. Serrano-Gotarredona, et al "CAVIAR: A 45k-Neuron, 5M-Synapse, 12G-connects/sec AER Hardware Sensory-Processing-Learning-Actuating System for High Speed Visual Object Recognition and Tracking," IEEE Trans. on Neural Networks, vol. 20, No. 9, pp. 1417-1438, September 2009.

[18] Jaer open source project. http://sourceforge.net/apps/trac/jaer/wiki

[19] T. Delbrück and P. Lichtsteiner, "Fast sensory motor control based on event-based hybrid neuromorphic-procedural system," Proc. of the IEEE Int. Symp. Circ. and Systems (ISCAS), pp. 845 - 848, 2007.

[20] P. Rogister, et al. "Asynchronous Event-based Binocular Stereo Matching". IEEE Trans. on Neural Networks and Learning Systems, Vol. 23, N. 2, pp. 347-353, 2012.

[21] J. Carneiro, S-H Ieng, C. Posch, and R. Benosman, "Asynchronous event-based 3D reconstruction from neuromorphic retinas," Neural Networks, vol. 45, pp. 27- 38, Sept. 2013.

[22] T. Serrano-Gotarredona, et al "Improved Contrast Sensitivity DVS and its Application to Event-Driven Stereo Vision," ISCAS 2013

[23] L. Camuñas-Mesa et al., "On the use of orientation filters for 3D reconstruction in event-driven stereo vision," Front. Neurosci. 8:48, 2104.