Improved Contrast Sensitivity DVS and its Application to Event-Driven Stereo Vision

T. Serrano-Gotarredona[1], J. Park[2], A. Linares-Barranco[3], A. Jiménez[3], R. Benosman[4], B. Linares-Barranco[1]

[1]Instituto de Microelectrónica de Sevilla-IMSE-CNM-CSIC, Spain
[2]University of California San Diego, USA
[3]Universidad de Sevilla, Spain
[4]Institut de Vision, UPMC, France
email: terese@imse-cnm.csic.es

Abstract- This paper presents a new DVS sensor with one order of magnitude improved contrast sensitivity over previous reported DVSs. This sensor has been applied to a bio-inspired event-based binocular system that performs 3D event-driven reconstruction of a scene. Events from two DVS sensors are matched by using precise timing information of their ocurrence. To improve matching reliability, satisfaction of epipolar geometry constraint is required, and simultaneously available information on the orientation is used as an additional matching constraint.

I. INTRODUCTION

Even the most modern artificial computing systems are known to underperform biological systems when dealing with intelligent processing of sensory signals. Tasks as object recognition in a visual scene, which are effortlessly done by human brains, take a long computation time in the most modern computers. It is obvious that the architecture and computation style in nervous systems is radically different from classical digital computing systems. Neuromorphic engineering is an interdisciplinary discipline that takes inspiration from biology to design artificial neural systems, such as vision systems, auditory processors, and autonomous robots, the physical architecture and design principles of which are based on those of biological nervous systems.

Address-Event-Representation (AER) sensors and processors are event-driven, that is, information is coded and transmitted as pulses (spikes or events), similarly to what happens in biological sensory and processing systems. In an AER sensor, each time a pixel senses relevant information it asynchronously sends an event out, which can be processed event by event by event-based processors [1]-[3]. That way, relevant features pass through all the processing levels almost instantaneously, the only delay being caused by the propagation of pulses along the processing chain. Because of their high speed potential and energy efficient nature, AER sensors have become very popular in recent years. A wide variety of AER vision sensors have recently appeared in literature [4]. Other non event-driven vision sensors coding output information as events but with frame-driven acquisition have also been reported [5]-[6].

Of special interest for very high speed processing applications are the so-called "Dynamic Vision Sensors" (DVS), where each pixel autonomously computes the normalized time derivative of the sensed light ($\dot{I}/I$) and provides an output event with its (x, y) coordinate when this amount exceeds a preset contrast [7]-[9]. DVS cameras have unique features such as contrast coding under very wide illumination variation, micro second latency response to fast stimuli, and low output data rate, which greatly improve the efficiency of post-processing stages. They can track extremely fast objects without special lighting conditions, providing timing resolutions better than 100kFrames/s. The availability of DVS cameras is triggering a new range of vision applications in the fields of surveillance, motion analyses, robotics, and microscopic dynamic observation.

In the real time reconstruction of three dimensional scenes, the high temporal resolution of current DVSs can open the way to new computation paradigms. Conventional stereovision methods are still computationally expensive due to the use of frame-based cameras. This is mainly due to the high amounts of redundancies conveyed by frames. Frame-based stereovision processes are also incompatible with precise timings usually used in the early visual areas of the brain [10]. Frame-based stereo rarely focuses on the link between images and scenes' dynamics. However, dynamic information is crucial for stereo matching as it introduces an additional temporal constraint in the recovery of scenes' structures [11]. Most of the conventional frame-based methods usually process sequentially sets of images regardless to what happened in the previous acquisitions. Frame-based stereo is also "token-based" meaning that direct matching of pixels is rarely performed, instead several features are used such as orientation [12], optical flow [13] but mainly corners-based descriptors of local luminance [14].

Asynchronous event-based acquisition allows a direct match of pixels based on the time of spiking. Stereo information can then be computed in a much faster manner with more relation to the way in which natural systems may process visual information. The principle is to derive stereo correspondence using the temporal occurrences of events. If two events happen at the same time and fulfill additional constraints linked to the pose of the cameras and scenes' content they are considered matched. Several studies focused on developing artificial neural computation of depth have been performed. The first event-based stereo vision system was implemented by Mahowald et al [15]. They implemented in a chip Marr-Poggio cooperative stereo algorithm [16] by detecting temporal coincidence of events coming from two spiking retinas using a correlator array. Another multi-chip stereo system uses a combination of AER Gabor-type chips and digital chips

This work has been funded by ERANET grant PRI-PIMCHI-2011-0768 (PNEUMA) funded by the Spanish Ministerio de Economia y Competitividad, Spanish research grant (with support from the European Regional Development Fund) TEC2009-10639-C04-01 (VULCANO), and Andalusian research grant TIC-6091 (NANONEURO). This work has been partially developed in the Capocaccia Cognitive Neuromorphic and Telluride Neuromorphic Cognition Engineering Workshops 2012

to implement disparity-tuned complex neurons constructed according to the binocular energy model [17]. Other neuromorphic stereo approaches use the output of the event-based DVS retinas in a classic frame-based approach to retrieve 3D information [18]. In [19], an event-based stereo tracking algorithm tracks the position of a moving object in both DVS retina views using an event-based mean-shift tracker and then reconstructs the position of the object in 3D.

In this paper, we present a new DVS sensor with better contrast sensitivity than previously reported ones [7]-[9] and demonstrate its application in event-based computation of 3D using the output from two DVS sensors connected to a convolution network hardware. This paper extends the work on precise timing to determine matches similar [17] by adding real-time orientation information to incoming events, thus increasing the reliability of matches. In Section II, we describe the DVS sensor used. The binocular system is described in Section III. Section IV concludes the paper.

II. THE DVS SENSOR

Fig. 1(a) shows the basic block diagram of a typical DVS pixel [7]-[8], [20]. The first stage transduces photo current to a voltage proportional to the logarithm of light

$$V_{log} = V_{DC} + A_v U_T \ln I_{ph} \quad (1)$$

where U_T is thermal voltage, A_v is a voltage gain factor, and V_{DC} is a light independent DC voltage level with high inter-pixel mismatch. The second stage amplifies V_{log} by C_1/C_2 resetting the charge integrated at C_2 every time $V_{diff}=(C_1/C_2)V_{log}$ varies a fixed threshold $\pm V_{th}$ set by the comparators. This also eliminates the DC component at V_{log}. The result is that each pixel generates a signed asynchronous output "event" every time its light changes by $\ln I_{ph}(t_2)-\ln I_{ph}(t_1) = \pm\theta_{ev}$, with $\theta_{ev} = C_2 V_{th}/(C_1 U_T A_v)$.

Consequently, pixel information is obtained not synchronously at fixed time steps δt (as in conventional video sensors), but asynchronously, driven by data at fixed relative light increments $\theta_{ev} = |\ln(I_{ph}(t_2)/I_{ph}(t_1))|$, as shown in Fig. 1(b). The DVS outputs thus focus on (relevant) information change while reducing data throughput. Parameter θ_{ev} represents the minimum detectable temporal contrast. Contrast sensitivity can consequently be minimized by maximizing overall voltage gain $A_T = A_v C_1/C_2$.

Fig. 1(c) shows the original photo transduction stage [7], [9] with $A_v = n_n$, where n_n is the subthreshold slope factor of NMOS transistor M_{n1}. A reasonable overall voltage gain was obtained by setting $C_1/C_2 = 20$, but this resulted in about 50% pixel area occupation by capacitors (unless a MiM process is used [9]). Fig. 1(d) shows Leñero's [8] photo transduction and pre-amplification stage, where $A_v \approx 12$ helped to reduce capacitive spread to $C_1/C_2 = 5$ while improving overall voltage gain to about $A_v C_1/C_2 \approx 60$ with smaller area. However, in Leñero's [8] scheme the pre-amplifier required to match NMOS and PMOS transistors and had high power consumption.

Fig. 1(e) illustrates the novel concept used for the DVS used in this work. Current mirror M_{p1}-M_{p2} amplifies photo current to $A_I I_{ph}$ [20] feeding a column of N diode-connected transistors M_{nj}, $j = 1,... N$. This transistor column performs a transimpedance amplification from input current $A_I I_{ph}$ to output voltage V_{log}. Assuming each diode-connected transistor is biased in weak inversion and using for its drain-to-source current I_{nj} the approximation

$$I_{nj} = A_I I_{ph} \approx I_{sn} e^{\frac{V_{Gj} - V_{Sj}}{n_n U_T}} \quad (2)$$

results in

$$V_{log} = \sum_{j=1}^{N} (V_{Gj} - V_{Sj}) = V_{DC} + N n_n U_T \log I_{ph} \quad (3)$$

with $V_{DC} = N n_n U_T \log(A_I/I_{sn})$. Consequently, the pre-amplification gain $A_v \approx n_n N$ is improved by a factor N. This improvement introduces no extra inter-pixel mismatch. Voltage headroom limits the practical number of stacked transistors to $N = 4$.

The prototype used in this work uses a chain of two low-power low-mismatch preamplification stages, achieving a maximum total pixel gain A_T about 120. The prototype has been fabricated in the double-poly 4-metal 0.35µm CMOS AMS technology, with the OPTO option. Fig. 2 shows a micro photograph of the fabricated prototype, including an inset with the pixel layout.The prototype exhibits a minimum achievable contrast sensitivity of 1.5% and a power consumption of 4mW. These figures have been improved by about one order of magnitude with respect to previous DVS designs [7]-[9]. A FPN down to 0.9% has been measured (25-50% lower with than previous designs)

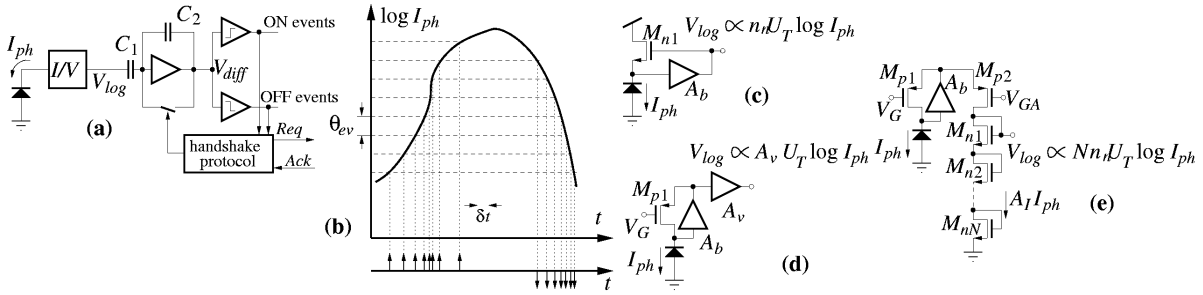


Fig. 1: (a) Pixel block diagram, (b) data driven asynchronous event generation, (c) Delbrück's original photo current transduction circuit, (d) Leñero's transduction with mismatch sensitive pre-amplification, (e) proposed transduction with mismatch insensitive pre-amplification.

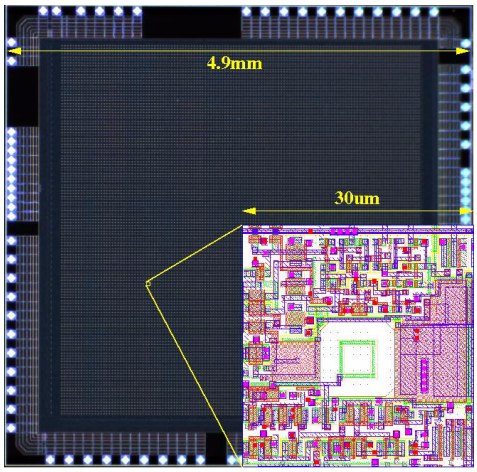


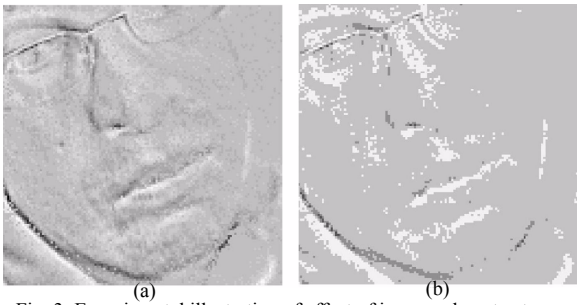
Fig. 2: Chip micro photograph and pixel layout

under a illumination of 60 lux, while measured latency was 3µs, kept equal to the best previously reported values. Maximum output event rate was kept high at 20*Meps* thanks to Boahen's burst-mode row parallel AER read out scheme [21]. The only penalty was that intra-scene illumination dynamic range was limited to about 3 decades (60dB). Table 1 summarizes the main design specifications.

Fig. 3 plots images captured with the present prototype illustrating the effect of improving contrast sensitivity. The images shown in Fig. 3 correspond to a moving face reconstructed by histogramming events during 30ms. Image in Fig. 3(a) was captured when setting the contrast sensitivity to 1.5%, while image in Fig. 3(b) corresponds to setting contrast sensitivity to about 10%.

Technology	0.35µm 4M 2P
Resolution	128x128
Chip Area	4.9x4.9mm^2
Pixel Area	30x31µm^2
Fill Factor	10.5%
Power Consumption (at 100keps)	4mW
Latency	3µs
Contrast Sensitivity	1.5%
FPN (%contrast)	0.9
Global Illumination DR	120dB
Intrascene Illumination DR	60dB
Max. Output Speed	20Meps

Table 1 Main Chip Specification



(a) (b)

Fig. 3: Experimental illustration of effect of improved contrast sensitivity in the present DVS when observing a moving face. (a) Reconstructed image when collecting during 30ms the events produced by the present DVS when contrast sensitivity is set to 1.5%. (b) Same scene, but now the sensor is set to a contrast sensitivity of about 10%.

III. THE BINOCULAR SYSTEM

Fig. 4 shows a photograph of the experimental set-up of the binocular system. The output of two DVSs sensors is connected to a merger board [22]. The merger board provides a perfect synchronization of the events coming from the two retinas, as they are arbitrated and sent out through a common AER bus where an additional address bit is added to identify the source retina that originated that event. The board is designed so that it can merge events coming out of up to four retinas. If a vision system with a larger number of retinas must be assembled it could be done by cascading several merger boards. The only limitation is the maximum event throughput that can be handled in a single bus and the size of the address bus.

In our binocular system, the output of the merger board is connected to a commercial Virtex6 prototyping board from Xilinx (ML605). This board can be programmed to hold a mesh of up to 8x8 real-time 64x64 pixel convolution arrays with programmable routers to configure the topology of the network of convolution filters [23]. The Virtex6 board was programmed to perform real-time edge extraction of the images sensed by the retina. Each 128x128 pixel retina event-flow was downsampled to 64x64 pixels and filtered by 3 Gabor filters programmed to detect edges oriented at -45º, 45º and 90º, respectively. Thus a total of 6 convolutions were programmed. The output flow of the Virtex6 board fused the 6 Gabor filter outputs together with the two retinas output. Each filter and retina copy was 64x64 pixels with rectified unipolar output (no sign bit), thus requiring only 12 bits. These eight 64x64 12-bit AER flows are fused into a single one of 192x192 pixels. Therefore, by using 16 bit events for the full output flow, we see the results as shown in Fig. 5(c).

Finally, a USBAERmini2 board [22] was used to timestamp all the events going out of the Virtex6 board and communicate the timestamped events to the computer through a high-speed USB2.0 port.

For the experimental evaluation of the 3D matching process, we placed the two retinas in front of a swinging cube, as can be seen in Fig. 5(a). Fig. 5(c) plots the

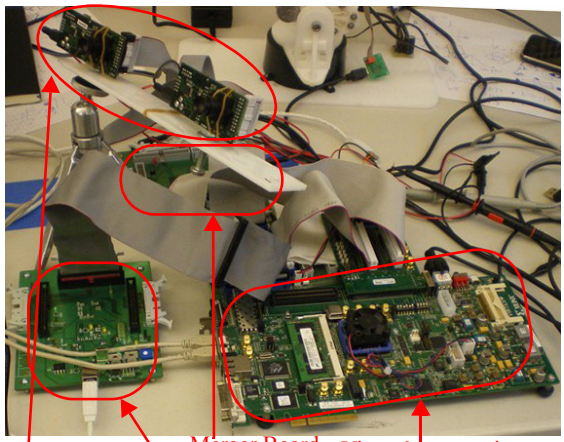
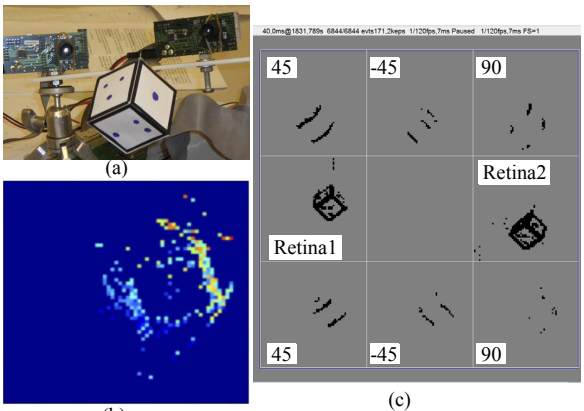


Fig. 4 Experimental set-up of the binocular system.



(a) (b) (c)

Fig. 5: Experimental set-up to compute depth. (a) Cube swinging in front of two retinas, (b) output of the 3D vision system with color coded depth, and (c) output of the Virtex6 board containing the output flow of the two retinas composed with the result of applying three gabor filters with different orientations to output of the retina 1 (lower row) and retina 2 (upper row), respectively.

composed output captured by the USBAERmini2 board, reconstructed during 40 ms. The Virtex6 board provides simultaneously the output flow of the two retinas and the results of applying three differently oriented Gabor filters to the output of each retina.

For the 3D matching process we use precise timing information of the events coming from the two retinas and increase the matching reliability by using edge orientation information as an additional constraint. That is, for each incoming event from the left retina, we compute its corresponding epipolar line in the right retina. An event of the right retina is considered to match if it arrives within a time window of 1 ms from the occurrence of the left event and it lies within its epipolar line and additionally its maximum response orientation corresponds to that of the left event. In that case, the two events are considered matched and a disparity value is computed. Fig. 5(b) shows the resulting event-driven disparity map computed for the swinging cube shown in Fig. 5(a). The disparity scale goes from dark blue to red to encode events from far to near.

IV. CONCLUSIONS

A new improved contrast sensitivity DVS sensor has been presented. The sensor has been applied in a neuromorphic 3D vision system. The system uses precise timing information between events coming from two DVSs to compute stereo matching. As a novel feature, this system computes real time gabor filter computation of the events coming out of the two retinas which are simultaneously available at the system output. This feature allows to improve the reliability of the matching process by using orientation as an additional matching constraint. More information and videos are available [24].

V. REFERENCES

[1] P. Venier, A. Mortara, X. Arreguit, and E. A. Vittoz, "An Integrated Cortical Layer for Orientation Enhancement," *IEEE J. Solid-State Circuits*, vol. 32, no. 2, pp. 177-186, Feb. 1997.

[2] T. Y. W. Choi, P. Merolla, J. Arthur, K. Boahen and B. E. Shi, "Neuromorphic Implementation of Orientation Hypercolumns," *IEEE Trans. on Circuits and Systems I*, vol 52, n. 6, pp. 1049-1060, June 2005.

[3] L. Camuñas-Mesa, C. Zamarreño-Ramos, A. Linares-Barranco, A. Acosta-Jiménez, T. Serrano-Gotarredona, and B. Linares-Barranco, "An Event-Driven Multi-Kernel Convolution Processor Module for Event-Driven Vision Sensors," *IEEE J. of Solid-State Circuits*, in Press, Feb. 2012.

[4] Tobi Delbrück, Bernabe Linares-Barranco, Eugenio Culurciello, Christoph Posch, "Activity-Driven, Event-Based Vision Sensors," *ISCAS* 2010, pp. 2426-29, May 2010.

[5] P. F. Ruedi, P. Heim, F. Kaess, E. Grenet, F. Heitger, P. Y. Burgi, S. Gyger, and P. Nussbaum, "A 128x128 pixel 120-dB dynamic-range vision sensor chip for image contrast and orientation extraction," *IEEE J. of Solid-State Circuits*, vol. 38, pp. 281-294, 2003.

[6] D. Kim and E. Culurciello, "A compact tri-mode vision sensor," *IEEE International Symposium on Circuits and Systems*, ISCAS 2010.

[7] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128x128 120 dB 15µs latency asynchronous temporal contrast vision sensor," *IEEE J. of Solid-State Circ.*, vol. 43, no. 2, pp. 566-576, Feb. 2008.

[8] J. A. Leñero-Bardallo, T. Serrano-Gotarredona and B. Linares-Barranco, "A 3.6µs latency asynchronous frame-free event-driven dynamic-vision-sensor," in *IEEE Journal of Solid-State Circuits*, vol. 46, No. 6, pp. 1443-55, June, 2011.

[9] C. Posch, D. Matolin, and R. Wohlgennant, "A QVGA 143dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," *IEEE J. of Solid-State Circuits*, vol. 46, no. 1, pp. 259-275, January 2011.

[10] M. Meister and M. J. B. Ii, "The neural code of the retina," *Neuron*, vol. 22, pp. 435–450, March 1999.

[11] P. Rogister, R. Benosman, S-H. Ieng, P. Lichtsteiner, T. Delbruck, "Asynchronous Event-based Binocular Stereo Matching". *IEEE Trans. on Neural Networks and Learning Systems*, Vol. 23, N. 2, pp. 347-353, 2012.

[12] G. Granlund and H. Knutsson, "Signal processing for computer vision," *Kluwer*, 1995.

[13] M. Gong, "Enforcing temporal consistency in real-time stereo estimation," *ECCV*, pp. 564–577, 2006.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints" *International Journal of Computer Vision,* 60, 2 (2004), pp. 91-110

[15] M. Mahowald and T. Delbrück, "Cooperative stereo matching using static and dynamic image features," in *Analog VLSI Implementation of Neural Systems*, C. M. Ismail and M., Eds. Boston: Kluwer Academic Publishers, 1989, pp. 213–238.

[16] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, no. 4262, pp. 283–287, 1976.

[17] E. K. C. Tsang and B. E. Shi, "A neuromorphic multi-chip model of a disparity selective complex cell," in *Advances in Neural Information Processing Systems*, vol. 16. MIT Press, 2004.

[18] J. Kogler, C. Sulzbachner, and W. Kubinger, "Bio-inspired stereo vision system with silicon retina imagers," *7th ICVS (Int. Conference on Computer Vision Systems)*, vol. 5815, pp. 174–183, 2009.

[19] T. Delbruck, "Dynamic Vision Sensor (DVS) - asynchronous temporal contrast silicon retina", http://siliconretina.ini.uzh.ch/wiki/index.php, 2009

[20] T. Serrano-Gotarredona, B. Linares-Barranco and A. G. Andreou, "Very wide range tunable CMOS/bipolar current mirrors with voltage clamped input," *IEEE Trans. on Circuits and Systems I,* pp. 1398-1407, Nov. 1999.

[21] K. Boahen, "Point-to-point connectivity between neuromorphic chips using address event," *IEEE Trans. on Circuits and Systems II*, vol. 47, pp. 416-434, May 2000.

[22] R. Serrano-Gotarredona et al., "CAVIAR: A 45k Neuron, 5M Synapse, 12GConnects/s AER Hardware Sensory–Processing–Learning–Actuating System for High-Speed Visual Object Recognition and Tracking," *IEEE Trans. on Neural Networks,* vol. 20, N. 9, pp. 1417- 1438, Sept. 2009.

[23] C. Zamarreño-Ramos, A. Linares-Barranco, T. Serrano-Gotarredona and B. Linares-Barranco, "Multi-casting mesh AER: a scalable assembly approach for reconfigurable neuromorphic structured AER systems. Application to ConvNets", *IEEE Trans. on Biomedical Circuit and Systems*, in Press.

[24] T. Serrano-Gotarredona, J. Park, A. Linares-Barranco, R. Benosman, and B. Linares-Barranco, "Convolution based event-driven stereo computation," available from http://neuromorphs.net/nm/wiki/learn12convolution