Event-driven stereo vision with orientation filters

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

S. Ieng and R. Benosman
Université de Pierre et Marie Curie
Institut de la Vision
Paris, FRANCE

Abstract— **The recently developed Dynamic Vision Sensors (DVS) sense dynamic visual information asynchronously and code it into trains of events with sub-micro second temporal resolution. This high temporal precision makes the output of these sensors especially suited for dynamic 3D visual reconstruction, by matching corresponding events generated by two different sensors in a stereo setup. This paper explores the use of Gabor filters to extract information about the orientation of the object edges that produce the events, applying the matching algorithm to the events generated by the Gabor filters and not to those produced by the DVS. This strategy provides more reliably matched pairs of events, improving the final 3D reconstruction.**

Keywords— Stereovision, Neuromorphic vision, Address Event Representation (AER), Event-driven processing, Convolutions, Gabor filters

I. INTRODUCTION

Biological vision systems are known to outperform any modern artificial vision technology. Traditional frame-based systems are based on capturing and processing sequences of still frames. This yields a very high redundant data throughput, imposing high computational demands. This limitation is overcome in bio-inspired event-based vision systems, where visual information is coded and transmitted as events (spikes). This way, much less redundant information is generated and processed, allowing for faster and more energy efficient systems.

Address Event Representation (AER) is a widely used bio-inspired event-driven technology for coding and transmitting (sensory) information [1]. In AER sensors, each time a pixel senses relevant information it asynchronously sends an event out, which can be processed by event-based processors [2]-[3]). This way, the most important features pass through all the processing levels very fast, as the only delay is caused by the propagation and computation of events along the processing network. Also, only pixels with relevant information send out events, reducing power and bandwidth consumption. These properties (high speed and low energy) are making AER vision sensors very popular.

The development of the Dynamic Vision Sensors (DVS) [4]-[6] was very important for high speed applications. These devices can track extremely fast objects with standard lighting conditions, providing a sampling rate higher than an equivalent of 100 KFrames/s. Exploiting this fine time resolution provides a new mean for achieving stereo vision with fast and efficient algorithms [7]. Frame-based methods usually process sequentially sets of images independently, searching for several features like orientation, optical flow or descriptors of local luminance [8]. However, event-based systems can compute stereo information much faster using the precise timing information to match pixels between different sensors. Other works implement neuromorphic disparity models without using explicitly spike timing [9]-[10].

In this paper, we explore different ways to improve the 3D object reconstruction using Gabor filters to extract orientation information from retina events. For that, we use two DVS sensors with 128 x 128 pixels and high contrast sensitivity (allowing the retina to detect contrast as low as 1.5%) [6], whose output is connected to a convolutional network hardware [3]. Different Gabor filter architectures are implemented to reconstruct the 3D shape of objects. Section II describes the calibration method used in this work. In Section III, we detail the matching algorithm applied, while Section IV provides experimental results. Section V concludes the paper.

II. STEREO CALIBRATION

Let us use lower case to denote a 2D point as $m = [x\ y]^T$, and capital letter to denote a 3D point as $M = [X\ Y\ Z]^T$. Augmented vectors are built by adding 1 as the last element: $\widetilde{m} = [x\ y\ 1]^T$ and $\widetilde{M} = [X\ Y\ Z\ 1]^T$. Under the assumptions of the pinhole camera model, the relationship between $\widetilde{m}$ and $\widetilde{M}$ is given by [11]:

$$\widetilde{m} = P_i \cdot \widetilde{M} \qquad (1)$$

where P_i is the projection matrix for camera i. Therefore, knowing the projection matrices of the different cameras in a vision system can be enough to extract the coordinates of the 3D points in space from their corresponding 2D projections.

The fundamental matrix F is a 3x3 matrix which relates the corresponding points between two cameras, and it can be defined by the equation:

$$\widetilde{m}_1^T F \widetilde{m}_2 = 0 \qquad (2)$$

where $\widetilde{m}_1$ and $\widetilde{m}_2$ are a pair of correspondent 2D points in both cameras [11].

In this work, we have implemented a calibration technique based on a known 3D object, consisting of 36 blinking LEDs distributed over two orthogonal planes (Fig.1). Using this fixed pattern, we calibrated two Dynamic Vision Sensors (DVS) that are only sensitive to changes in luminance [6], solving (1) and (2) as shown in [12].

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). The authors also benefited from both the CapoCaccia Cognitive Neuromorphic Engineering Workshop, Sardinia, Italy, and the Telluride Neuromorphic Cognition Engineering Workshop, Telluride, Colorado.

978-1-4799-3432-4/14/$31.00 ©2014 IEEE

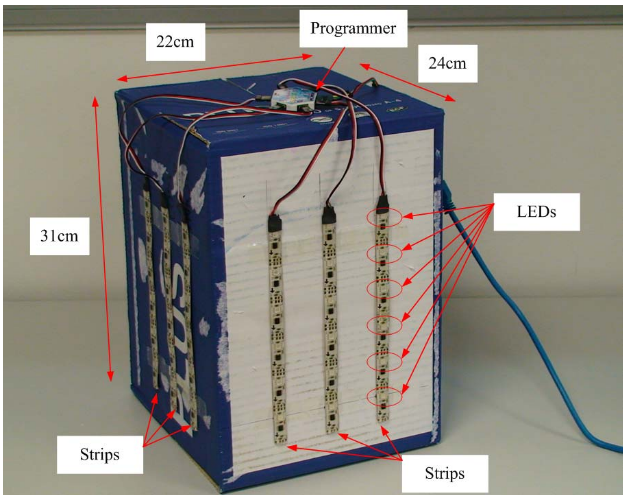

Fig. 1. Calibration object: 36 LEDs distributed over two orthogonal planes.

III. EVENT MATCHING

In stereo vision systems, a 3D point in space M is projected onto the focal planes of both cameras in pixels m_1 and m_2, therefore generating events $e(m_1^i, t)$ and $e(m_2^i, t)$. Reconstructing the original 3D point requires matching each pair of events produced by point M at time t [13]. For that, we implemented a matching algorithm based on a list of constraints applied to each event.

A. Temporal match

One of the most useful advantages of event-driven DVS based vision sensing and processing is the high temporal resolution down to fractions of micro seconds [4], [5], [6]. Thus, in theory, two identical DVS cameras observing the same scene should produce corresponding events simultaneously [7]. However, in practice, there are many non-ideal effects that end up introducing appreciable time differences (up to many milli seconds) between corresponding events [6]. Nonetheless, corresponding events occur within a milli second range time window, depending on ambient light (the lower light, the wider the time window). As a consequence, this first restriction implies that for an event $e(m_1^i, t_1)$, only those events $e(m_2^i, t_2)$ with $|t_1 - t_2| < \delta_t/2$ can be candidates to match.

B. Epipolar restriction

As is described in detail in [11], when a 3D point in space M is projected onto pixel m_1 in retina 1, the corresponding pixel m_2 lies on a specific line in retina 2 called the "epipolar line" [13]. Using this property, a second restriction is added to the matching algorithm using the fundamental matrix F to calculate the epipolar line Ep_2 in retina 2 corresponding to event m_1 in retina 1 ($Ep_2(m_1) = F^T \widetilde{m}_1$). Therefore, only those events $e(m_2^i, t_2)$ whose distance to Ep_2 is less than a given limit δ_{Ep_i} can be candidates to match.

C. Ordering constraint

For a practical stereo configuration of retinas where the angle between their orientations is small enough, a certain geometrical constraint can be applied to each pair of corresponding events. In general, the horizontal coordinate of the events generated by a retina is always larger than the horizontal coordinate of the corresponding events generated by the other retina.

D. Polarity

The silicon retinas used in our experimental setup generate output events when they detect a change in luminance in a pixel, indicating in the polarity of the event if that change means increasing or decreasing luminance [4]-[6]. Therefore, we can impose the condition that two corresponding events in both retinas must have the same polarity.

E. Orientation

If the focal planes of two retinas in a stereo vision system are vertically aligned and have a small horizontal vergence, the orientation of observed edges will be approximately equal. A static DVS produces events when observing the edges of moving objects. Therefore, correspondent events in the two retinas are produced by the same moving edges, and consequently the observed orientation of the edge should be similar in both retinas.

The application of banks of Gabor filters to the events generated by both retinas provides information about the orientation of the object edges that produce the events. This way, by using Gabor filters with different angles we can apply the matching algorithm to the oriented events produced by the filters (instead of the original events produced by the DVS), so that only events with the same orientation can be matched. Thus, the events coming out of retinas R_1 and R_2 are processed by Gabor filters G_{1x} and G_{2x}, respectively (with $x = 1, 2, \dots N$, being N the number of filters for each retina). Then, for each pair of Gabor filters G_{1x} and G_{2x}, conditions A to D are applied to obtain matched events for each orientation. Therefore, the final list of matched events will be obtained as the addition of all the lists of events obtained for each orientation.

IV. RESULTS

In this section, we describe briefly the hardware setup used for the experiments, we validate the calibration method and finally we present results on the reconstruction of 3D objects.

A. Hardware setup

The event-based stereo vision processing technique outlined above has been tested using two DVS sensor chips with 128 x 128 pixels [6] whose outputs are connected to a merger board [14] which sends the events to a 2D grid array of event-based convolution modules implemented within a Spartan6 FPGA [3]. The Spartan6 was programmed to perform real-time edge extraction on the visual flow using Gabor filters, as described in detail in [2]. Finally, a USBAERmini2 board [14] was used to timestamp all the events going out of the Spartan6 board and send them to a computer through a high-speed USB2.0 port (Fig.2).

Each convolution module in the FPGA consists of 128 x 128 pixels which behave like leaky integrate-and-fire neurons. Several convolutional modules can be arranged in a 2D mesh, each one communicating bidirectionally with all four neighbors. Each module is characterized by its coordinate within the array. Address events are augmented by adding either the source or destination module coordinate. Each module includes an AER router which decides how to route

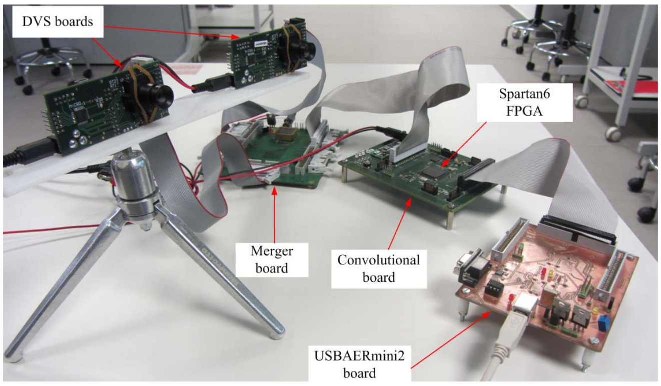

Fig. 2. Experimental stereo setup.

the events [3]. This way, any network architecture can be implemented by programming the appropriate kernel in each convolutional module to extract a specific orientation.

B. Calibration results

In order to calibrate the stereo setup, we built a structure of 36 blinking LEDs distributed over two orthogonal planes (Fig.1). This structure was placed in front of the DVS sensors at approximately $1m$ distance, and the events generated by the retinas were recorded in the computer. This recording was processed offline to obtain the 2D coordinates of the LEDs projected in both retinas. For that, we represented a 2D image coding the number of spikes generated by each pixel. All those pixels with a number of spikes below a certain threshold were set to zero, while all those pixels above the threshold were set to one, obtaining a binarization of the image. Then, for each cluster of pixels we calculated the mean coordinate, obtaining the 2D projection of the LEDs with sub-pixel resolution.

These 2D coordinates together with the known 3D positions of the LEDs in space were used to calculate the projection matrices P_1 and P_2, and the fundamental matrix F following the methods described in Section II. To validate the calibration, P_1 and P_2 were used to reconstruct the 3D calibration pattern. The reconstruction error is measured as the distance between each original 3D point and its reconstructed position. The mean reconstruction error obtained is $2mm$ with a standard deviation of $1mm$. This error is comparable to the size of each LED.

C. 3D reconstruction

For the experimental evaluation of the 3D reconstruction, we analyzed the effect of several configurations of Gabor filters on the event matching algorithm, using different numbers of orientations (from 2 to 8) and 4 different spatial scales of kernels. In general, the different angles implemented in each case are uniformly distributed between 90° and -90°. These configurations are compared with the implementation of the matching algorithm without Gabor filters (i.e. without the orientation constraint).

A swinging pen of 14cm was placed in front of the two retinas for half a minute, with a number of approximately 100 Kevents generated by each retina. Fig.3(a) shows the number of events processed by the algorithm (which correspond to the outputs of the Gabor filters) for different numbers of orientations and scales. The horizontal line represents the number of events when no Gabor filters are applied. We show only the number of events coming originally from Retina 1, as they both have been configured to generate approximately the same number of events for a given stimulus.

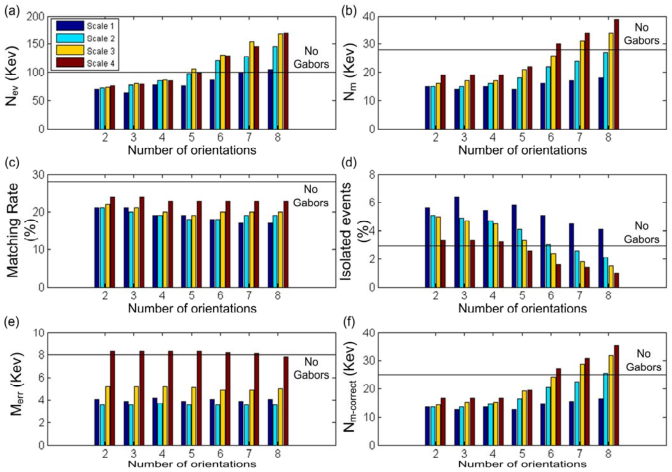

Fig. 3. Comparison of the 3D reconstruction results. Vertical bars represent the results obtained applying Gabor filters with different number of orientations and scales, while horizontal lines indicate the values obtained without Gabor filters. (a) Total number of events processed by the matching algorithm. (b) Number of matched events produced by the algorithm. (c) Ratio of matched events over the total number of events. (d) Ratio of isolated events over the total number of matched events. (e) Ratio of wrongly matched events over the total number of matched events. (f) Number of correctly matched events.

When the algorithm is applied directly to the output of the retinas, the number of matched pairs of events obtained is around 28 Kevents (28% of success rate). Fig.3(b) shows the number of matched events for each configuration of Gabors. If we calculate the percentage of success obtained by the algorithm for each configuration, we obtain the values shown in Fig.3(c).

Although these results show that the matching rate of the algorithm is smaller when we use Gabor filters to extract information about the orientation of the edges that generated the events, we should consider that the performance of 3D reconstruction is determined by the total number of matched events, not the relative proportion. For that reason, we consider that a bank of 8 Gabor filters with kernels of scale 4 gives the best result, with more than 39 Kevents that can be used to reconstruct the 3D sequence, using 100 Kevents generated by the retinas.

Another parameter that can be used to measure the quality of the 3D reconstruction is the proportion of "isolated" events in the matched sequence. An event is considered as isolated when it is not correlated to any other event in a certain spatio-temporal window. With this definition, the 28 Kevents that were matched for the retinas without any Gabor filtering were used to reconstruct the 3D coordinates, resulting in 2.93% of noise events. After the application of the same methodology to all the Gabor filter configurations, the results in Fig.3(d) are obtained. These results show that several configurations of Gabor filters give a smaller proportion of isolated events.

Calculating continuously the mean and standard deviation of the distribution of disparities, we define the range of acceptable values, and we identify as wrongly matched all those events whose disparity is outside that range. Using this

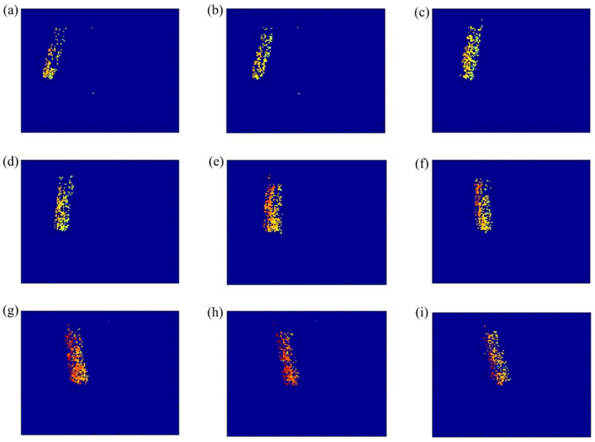

Fig. 4. Sequence of disparity maps reconstructed with $T_{frame} = 50ms$ corresponding to the movement of the swinging pen (from (a) to (i)). The disparity scale goes from dark blue to red to encode events from far to near.

method, we calculate the proportion of wrongly matched events and present it in Fig.3(e). Finally, Fig.3(f) shows the number correctly matched events, subtracting both the isolated and wrongly matched events from the total number of matched events.

Using the sequence of matched events provided by the algorithm in the best case (8 orientations, scale 4, which gives the largest number of correctly matched events), we computed the disparity map. For that, we calculated the euclidean distance between both pixels in each pair of events. This measurement is inversely proportional to the distance between the represented object and the retinas, as further objects produce a small disparity and closer objects produce a large disparity. Fig. 4 shows 9 consecutive frames of the obtained disparity sequence, with a frame time of $50ms$.

Finally, the 3 dimensional coordinates of the matched events are calculated using P_1 and P_2. Fig. 5 shows 9 consecutive frames of the resultant 3D reconstruction video, with a frame time of $50ms$. The shape of the pen is clearly represented as it moves around 3D space. Using these events, we measured manually the approximate length of the pen, obtaining 14.85cm, which means an error of 0.85cm.

V. CONCLUSIONS

Event matching algorithms have been proposed for stereo reconstruction, taking advantage of the precise timing information provided by DVS sensors. These algorithms are based on a set of constraints that must be fulfilled by corresponding events. In this work, we have explored the benefits of using Gabor filters to extract the orientation of the object edges in order to apply the matching algorithm to the oriented events, and not to those produced by the DVS. By analyzing different numbers of filters with several spatial scales, we have shown that we can increase the number of reconstructed events for a given input recording, reducing the number of isolated and wrongly matched events at the same time. In particular, we have shown an increase of up to 44% in the number of correctly matched events.

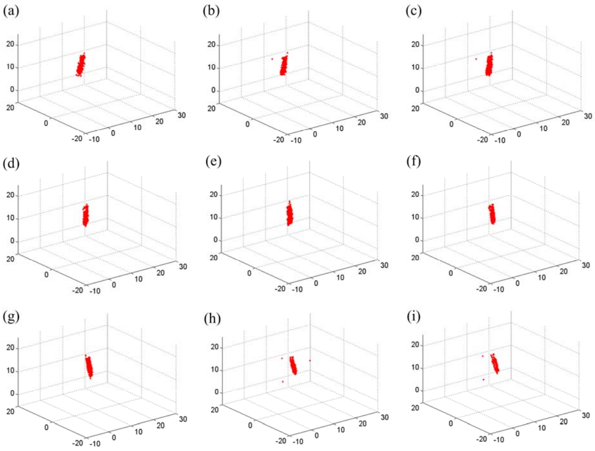

Fig. 5. 3D reconstruction of the swinging pen. Each plot corresponds to a $50ms$-frame representation of the 3D coordinates of the matched events.

REFERENCES

[1] M. Sivilotti, "Wiring considerations in analog VLSI systems with application to field-programmable networks," PhD dissertation, California Institute of Technology, Pasadena, CA, 1991.

[2] L. Camuñas-Mesa et al., "An event-driven multi-kernel convolution processor module for event-driven visión sensors," IEEE Journal on Solid-State Circuits, 47 (2): 504 – 517, 2012.

[3] 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 Transactions on Biomedical Circuits and Systems, 7 (1): 82 – 102, 2013.

[4] P. Lichtsteiner, C. Posch and T. Delbrück, "A 128x128 120dB 15μs latency asynchronous temporal contrast vision sensor," IEEE Journal on Solid-State Circuits, 43 (2): 566 – 576, 2008.

[5] C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," IEEE Journal on Solid-State Circuits, 46 (1): 259 – 275, 2011.

[6] T. Serrano-Gotarredona, and B. Linares-Barranco, "A 128x128 1.5% contrast sensitivity 0.9% FPN 3μs latency 4mW asynchronous frame-free dynamic vision sensor using transimpedance amplifiers," IEEE Journal on Solid-State Circuits, 48 (3): 827 – 838, 2013.

[7] P. Rogister, R. Benosman, S. Ieng, P. Lichsteiner, and T. Delbruck, "Asynchronous event-based binocular stereo matching," IEEE Transactions on Neural Networks, 23: 347 – 353, 2012.

[8] D. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60 (2): 91 – 110, 2004.

[9] E. K. C. Tsang, and B. E. Shi, "A preference for phase-based dispity in a neuromorphic implementation of the binocular energy model", Neural Computation, 16 (8): 1579 – 1600.

[10] E. K. C. Tsang, S. Y. M. Lam, Y. Meng and B. E. Shi, "Neuromorphic implementation of active gaze and vergence control", ISCAS 2008: 1076 – 1079.

[11] R. Hartley, and A. Zisserman, "Multiple View Geometry in computer vision," Cambridge University Press, 2003.

[12] R. Benosman, S. Ieng, P. Rogister, and C. Posch, "Asynchronous event-based Hebbian epipolar geometry," IEEE Transactions on Neural Networks, 22 (11): 1723 – 1734, 2011.

[13] J. Carneiro, S. Ieng, C. Posch, and R. Benosman, "Asynchronous event-based 3D reconstruction from neuromorphic retinas," Neural Networks, 45: 27 – 38, 2013.

[14] 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, 20 (9): 1417 – 1438, 2009.