Neural Process Lett (2016) 43:311–326
DOI 10.1007/s11063-015-9434-5

Asynchronous Event-based Cooperative Stereo Matching Using Neuromorphic Silicon Retinas

Mohsen Firouzi[1,2,3] · **Jörg Conradt**[1,2,3]

Published online: 26 May 2015

Abstract Biologically-inspired event-driven silicon retinas, so called dynamic vision sensors (DVS), allow efficient solutions for various visual perception tasks, e.g. surveillance, tracking, or motion detection. Similar to retinal photoreceptors, any perceived light intensity change in the DVS generates an event at the corresponding pixel. The DVS thereby emits a stream of spatiotemporal events to encode visually perceived objects that in contrast to conventional frame-based cameras, is largely free of redundant background information. The DVS offers multiple additional advantages, but requires the development of radically new asynchronous, event-based information processing algorithms. In this paper we present a fully event-based disparity matching algorithm for reliable 3D depth perception using a dynamic cooperative neural network. The interaction between cooperative cells applies cross-disparity uniqueness-constraints and within-disparity continuity-constraints, to asynchronously extract disparity for each new event, without any need of buffering individual events. We have investigated the algorithm's performance in several experiments; our results demonstrate smooth disparity maps computed in a purely event-based manner, even in the scenes with temporally-overlapping stimuli.

Keywords Silicon retina · Event-based stereo matching · Cooperative network · Frameless 3D vision · Disparity detection

✉ Mohsen Firouzi
mohsen.firouzi@tum.de
http://www.nst.ei.tum.de

Jörg Conradt
conradt@tum.de

[1] Neuroscientific System Theory, Technische Universität München, Karlstr. 45, 80333 Munich, Germany

[2] Bernstein Center for Computational Neuroscience, Munich, Germany

[3] Graduate School of Systemic Neurosciences, Ludwig-Maximilian University of Munich, Munich, Germany

1 Introduction

Depth perception is a crucial skill of animals and humans for survival. A predator is able to catch the prey in a very fast time scale, cats can jump onto the table, and birds can land on narrow edges and catch insects at the first attempt. These abilities and in general all behaviors that involve moving around in an environment require a precise estimate of how far away a visual object might be. In general there are two major types of cues in the environment that help animals in depth perception: external cues that are captured just by using one eye (monocular cues), e.g. perspective or relative size of the objects in the scene; and internal cues that rely on the physiological processing of the visual stimuli. Neurons in the visual cortex can compute distance using motion parallax cues and relative movement of retinal images [1]. The most important internal cue is retinal disparity, which is defined as the difference between positions of the objects on the retinal images. This cue is the anatomical consequence of the eyes' positions on the face. The ability to use retinal disparity in depth perception is known as stereopsis in vision and still is an active research field.

Following the fact that many basic aspects of the human visual processing system have been discovered in recent years, VLSI technology addresses emulating brain circuitry by introducing new microchips and brain-like information processing circuits, e.g. dynamic vision sensors (DVS), Silicon cochlear, and massively distributed processors (SpiNNaker) [2–4]. One open question is how to introduce neurophysiologically plausible stereopsis into technical systems by engineering underlying algorithmic principles of stereo-vision. The first attempt to answer this question was performed by David Marr who proposed a laminar network of sharply tuned disparity detector neurons, called cooperative network, to algorithmically model basic principles of the disparity detection mechanism of the brain [5]. In 1989 Mahowald and Delbruck developed a micro circuit which was the first hardware implementation of Marr's cooperative network [6].

1.1 Classic Stereo Vision Problem

Besides the important role of 3D sensing in living systems, adding depth information into 2D visual information enables artificial systems, e.g. robots and assistive devices to operate in the environment with more reliability. Generally in classic stereo vision, two cameras are used which are mounted on a common baseline to capture the scene from two different view-points. Geometrical characteristics of the stereo cameras can be formulated to map a single pixel of one image into a set of possible corresponding pixels in the other image [7]. Finding the matching objects or features in each stereo images is called "correspondence problem". There are two general classic algorithms to solve the correspondence problem; area-based and feature-based matching. In area-based algorithms, usually the intensity of an area around a single pixel is individually compared with a window around potential corresponding pixels in the other image. In feature-based matching, the correlation amongst features in each image is analyzed rather than intensity of pixels [8]. Classical stereo-matching algorithms are computationally demanding, since they require processing a stream of frames that often contains redundant background information. This problem impedes applicability of classic algorithms in applications in which the processing time is crucial, e.g. driving assistive devices and motion analysis [9].

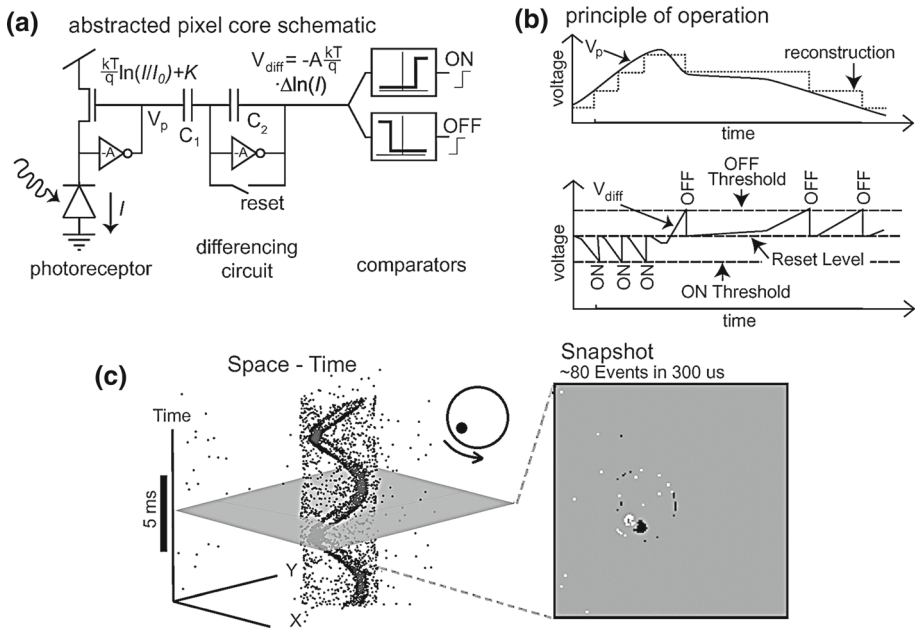

Fig. 1 **a** Abstracted circuitry of single pixel in DVS; **b** principle operation of single DVS pixel; **c** space-time representation of the event stream generated by a spinning disk, and single snapshot of the event stream in 300 μs (taken from [10])

1.2 Stereo Vision Using Silicon Retina Technology

Dynamic vision sensors that mimic basic characteristics of human visual processing, have created a new paradigm in vision research [10]. Similar to photoreceptors in the human retina, a single DVS pixel (receptor) can generate spikes (events) in response to a change of detected illumination. Events encode dynamic features of the scene, e.g. moving objects, using a spatiotemporal set of spikes (see Fig. 1c). Since DVS sensors drastically reduce redundant pixels (e.g. static background features) and encode objects in a frame-less fashion with high temporal resolution (about 1 μs), it is well suited for fast motion analyses, tracking and surveillance [11–15]. These Sensors are capable of operating in uncontrolled environments with varying lighting conditions because of their high dynamic range of operation (120 dB).

Although DVS sensors offer some distinguished capabilities, developing event-based processing algorithms and particularly stereo matching, is considered as a big challenge in literature [11,16–19]. The fact that conventional frame-based visual processing algorithms cannot fully utilize main advantages of DVS necessitates developing efficient and sophisticated event-driven algorithms for DVS sensors. The main line of research in event-based stereo matching using DVS is focused on temporal matching [16,17]. Kogler et al. [16] proposed a purely event-driven matching using temporal correlation and polarity correlation of the events. Due to intrinsic jitter delay and latency in a pixel's response which varies pixel by pixel [17], temporal coincidence alone is not reliable enough for event matching especially when the stimuli generate temporally-overlapping stream of events (i.e. when multiple different objects are moving in front of the cameras). Rogister et al. [17] combined epipolar constraint with temporal matching and ordering constraints to eliminate mismatched events

and have demonstrated that additional constraints can enhance the matching quality. Despite the fact that this method can partly deal with temporally-overlapping events, it still requires event-buffering for a time frame. To reduce ambiguity during the matching process, Carneiro et al. [18] have shown that by adding additional cameras to the stereo setup (Trinocular vision vs. Binocular vision), it is possible to find unique corresponding matching event pairs using temporal and epipolar constraints. The results in this work show a significant enhancement in the quality of event-based 3D-reconstruction compared with other methods though Trinocular vision is not biologically realistic. Taking into account the epipolar constraint, it is necessary to calibrate cameras and to drive algebraic equations of 3D geometry [7]. Therefore adding more cameras to the stereo setup will increase the complexity of the geometrical equations in spite of reducing ambiguity.

In this paper we propose a new fully event-driven stereoscopic fusion algorithm using silicon retinas. Event-driven matching means: as a single event occurs (caused by any motions or contrast changes in the scene), the algorithm should deal with it immediately and asynchronously without any need to collect events and construct a single frame. The main idea of our algorithm is borrowed from David Marr's cooperative computing approach [5]. Marr's cooperative network can just operate on the static features to deal with the correspondence problem. In this work we have formulated a dynamic cooperative network in order to take into account temporal aspects of the stereo-events in addition to existing physical constraints such as cross-disparity uniqueness and within-disparity smoothness. The network's input includes the retinal location of a single event (pixel coordinates) and the time at which it has been detected. Then according to the network's internal state (activity of the cells), which is shaped by previously fused events, disparity is extracted through a cooperative mechanism. The extracted disparity values can be further used for depth calculation of the events. The pattern of interaction amongst cells (suppression or excitation) applies physical and temporal constraints. In Chapter 4 we evaluated the proposed algorithm in several real-world experiments and the results demonstrate the accuracy of the network even with temporally-overlapping stimuli.

In the next section we will briefly describe the basic functionality of the Silicon Retina and in chapter 3 our event-based algorithm is elaborated in detail. In chapter 4 experimental results are shown and finally, the conclusion and remarks are presented in chapter 5.

2 Neuromorphic Silicon Retina

In 1991 Mahowald and Mead developed the first silicon retina to bring principle functionality of the human retina into VLSI circuits [20]. The basic operation of today's DVS sensors is similar to Mahowald and Mead's silicon retina whose pixels consist of a single CMOS photoreceptor to detect light intensity, differencing circuitry to compute change of the contrast or equivalently illumination, and comparator circuit to generate output spikes. Fig. 1a shows basic schematic of a single DVS pixel, where the light intensity is detected by a photoreceptor in the form of a current signal I and the current signal is amplified and transformed into a voltage signal V_p. The differencing circuit generates V_{diff} signal, which is proportional to change of log intensity ($V_{diff} \propto \Delta \ln(I)$). Finally, the comparator circuit compares the change of log intensity with preset thresholds. Therefore, if V_{diff} exceeds one of the ON or OFF thresholds (see Fig. 1b), the sensor will signal out a single event. Each event consists of a data packet including pixel coordinates and the type of the event (ON and OFF events for positive and negative intensity change respectively). Finally, activated pixel will be reset into

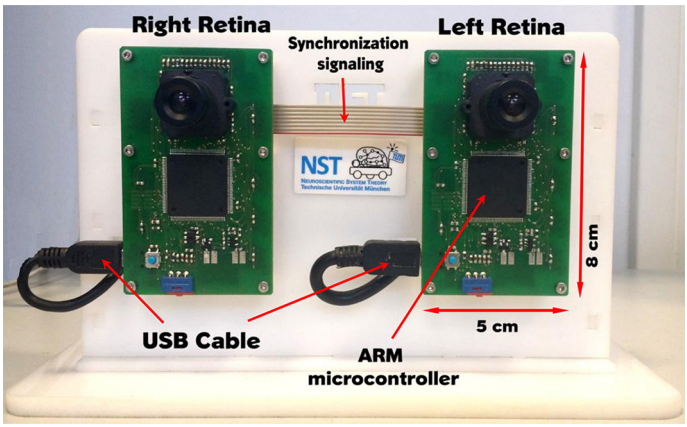

Fig. 2 Synchronized embedded stereo-DVS, eDVS4337

the base voltage (Fig. 1b). The encoding mechanism of the light using log intensity allows the sensor to operate in a wide range of illumination [10]. Figure 1c shows the space-time representation of an event stream for a rotating disk, and a single snapshot of the events within 0.3 ms. The type of the events are indicated by dark and white pixels.

It is worth to notice that each pixel in DVS is independent from other pixels and the data communication is asynchronous. This means that events are transmitted only once after they occur without a fixed frame rate and are independent from each other. The sensor chip we use in this work is the *DVS128* sensor with 128×128 spatial resolution, 1 μs temporal resolution and 120 dB dynamic range of operation [10].

2.1 Stereo Dynamic Vision Sensor

To fetch events and to stamp them with the time they have been generated, a supplementary circuit is required. In this work we use eDVS4337 circuit as a light, compact and low power solution that is used in several robotic applications and anomaly detection [15,21–24]. This embedded system uses an LPC4337 ARM Cortex microcontroller to fetch events from the retina and to control the data path and communication links. The stereo setup is built using two eDVS mounted side-by-side so that silicon retinas are 10 cm apart (Fig. 2). To synchronize two embedded boards, a Master/Slave signaling mechanism is performed on two microcontrollers before event fetching. Two separate USB links are connected to an external PC to read out event packets (Fig. 2). Each packet consists of the retinal position of the event (x coordinate and y coordinate), the time-stamp t which is created by microcontrollers, and the type of the event which is called polarity p.

3 Event-based Cooperative Stereo Matching

3.1 Cooperative Computing

Cooperative computing refers to the algorithms with distributed local computing elements that are interacting with each other using local operations. These operators apply specific constraints to the inputs in order to obtain a global organization and to solve a problem. The

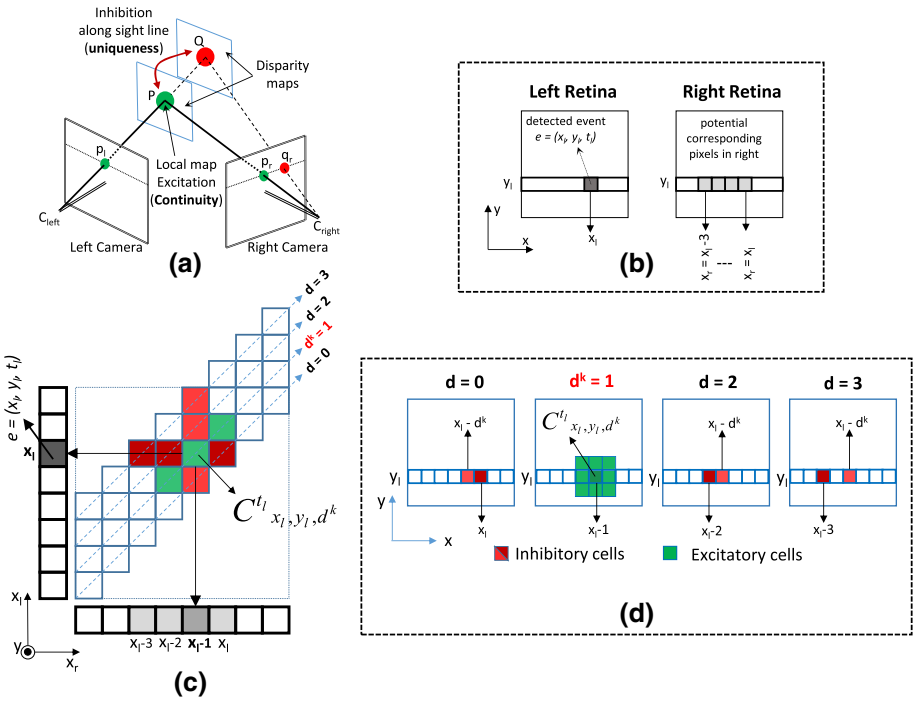

Fig. 3 **a** Topological description of continuity and uniqueness constraints in stereopsis problem **b** single detected event $e = (x_l, y_l, t_l)$ in left retina (*dark grey*) and potential corresponding candidates in right retina (*light grey*). **c** Cross-section scheme of the disparity maps, and stereo-retinas for $y = y_l$ **d** 2D disparity maps; the excitatory and inhibitory cells for $C_{xl,yl,dk}$ are indicated by *red* and *green* respectively. (Color figure online)

dynamics of these algorithms should reach a stable point given the inputs, to guarantee a unique solution for the problem they model. The pattern of interaction or equally the local operators, should be derived to computationally enforce the constraints amongst inputs. David Marr was the first who proposed a cooperative network to address the stereo matching problem. This network is composed of a matrix whose cells are created by the intersection of the pixel pairs in the left and the right images. Each single cell encodes the internal local belief in matching the associated pixel pairs. To derive the connectivity pattern between cells, two general physical constraints must be considered:

- *Cross-disparity uniqueness* reinforces a single pixel to possess one unique disparity value. Equivalently it means there must be a single unique corresponding pixel pair in each stereo images (in the case of no occlusion). So the cells that lie along the line-of-sight must inhibit each other to suppress false matching. For instance, in Fig. 3a, given p_l on the left image as a retinal projection of the object P, there are two matches in the right retina, p_r or q_r. But since just one of the candidates can be chosen, the cells that show the belief of $p_r - p_l$ correspondence i.e. P in Fig. 3a, should inhibit other cells that lie along disparity maps i.e. Q in Fig. 3a.
- *Within-disparity continuity* Since physical objects are cohesive, the surface of an object is usually smooth and should be emerged by a smooth disparity map. Therefore, neighboring cells that are tuned for a single disparity or equivalently lie in a common disparity map, should potentiate each other to generate a spatially smooth disparity map (see Fig. 3a).

In spite of many unanswered questions about neurophysiological mechanisms of the disparity detection, nowadays it is widely accepted that mammalian brains utilize a competitive process over disparity sensitive populations of neurons, to encode and detect horizontal disparity [25]. Similarly in the cooperative network, the cells are sharply tuned for a single disparity value. Furthermore, the pattern of suppression and potentiation has implemented a competitive mechanism in order to remove false matching and to reach a global solution. Basically, the standard cooperative dynamics can extract spatial correlation of the static features in the scene, but the question arises how to formulate and to construct an event-driven cooperation process to deal with dynamic features, e.g. DVS event stream. In the following section we will address this question in detail.

3.2 Event-driven Cooperative Network: Architecture and Dynamics

Given event $e = (P_l, t_l)$ detected in the left retina at time t_l and $P_l = (x_l, y_l)$ as pixel coordinates (dark grey in Fig. 2b), the set of possible corresponding pixels in the right retina will be as follows:

$$S_e = \left\{(x_r, y_r) \,\middle|\, x_l - d^{\max} \leq x_r \leq x_l \ , \quad y_r = y_l\right\} \tag{1}$$

where d^{max} is an algorithmic parameter that determines the maximum detectable disparity. For each possible matching pixel pair $P_l = (x_l, y_l)$ and $P_r = (x_r, y_r) \epsilon S_e$, a single cell $C_{xl,yl,dk}$ is created such that its temporal activity indicates the correspondence of that pixel pair. For instance, for the left event e in Fig. 3b, cell $C_{xl,yl,dk}$ is created as the intersection of x_l in the left and $x_r = x_l - d^k$ in the right images. $d^k = 1$ shows $C_{xl,yl,dk}$ is sensitive to disparity "1" (see Fig. 3c). So the network consists of a 3D matrix of the cells. The size of the matrix is $N \times N \times \mathrm{d}^{\max}$, in which N is the sensor resolution (Fig. 3c). If we expand the diagonal elements of the cooperative matrix into 2D maps, we can see a set of 2D disparity maps whose cells are specialized to detect one single disparity value (Fig. 3d). The cross section of the disparity maps for $y = y_l$ is shown in Fig. 3c.

Since the stereo retinal images are aligned with each other and lie in a common baseline, in Eq. (1) we assume epipolar lines are parallel with x axis and potential corresponding pixels must have a common y coordinate. This assumption is true when stereo cameras are placed on a same surface and in parallel to each other [7]. To apply physical and temporal constraints on the input events, we should properly formulate the pattern of interaction among the cells. In order to support the continuity constraint and similar to the classic cooperative network, the neighboring cells lying on a common disparity map should potentiate each other creating a local pattern of excitation. The set of excitatory cells for a single cell $C_{x,y,dk}$ is indicated by green color in Fig. 3c, d and can be described by following equation:

$$E(C_{x,y,d^k}) = \left\{C_{x',y',d^k} \,\middle|\, |x - x'| \leq r, |y - y'| \leq r\right\} \tag{2}$$

Having a unique possible matching pixel pair, the cells which lie along the line-of-sight should inhibit each other. So accordingly there are two patterns of inhibition:

- The first set of inhibitory cells which is shown in Fig. 3d by dark red, includes the cells that are topologically created by the intersection of the left pixel $P_l = (x_l, y_l)$, and all possible candidate pixels in the right ($P_r \epsilon S_e$):

$$I_1(C_{x,y,d^k}) = \left\{C_{x',y',d} \,\middle|\, 0 \leq d \leq d^{\max}, d \neq d^k, x' = x - d, y' = y\right\} \tag{3}$$

- A single candidate pixel in the right image (e.g. $x_r = x_l - 1$ in Fig. 3c), may have been chosen as a matching pixel for a former left event. Thus, a second set of inhibitory cells

are selected in order to suppress this group of false matching (indicated by light red in Fig. 2c, d).

$$I_2(C_{x,y,d^k}) = \left\{ C_{x',y',d} \,\middle|\, 0 \le d \le d^{\max}, \quad d \ne d^k \ , \ x' = x - d^k, \ y' = y \right\} \tag{4}$$

Descriptive features in DVS sensors are dynamic asynchronously-generated streams of events. Despite the fact that event matching based on exact temporal coincidence is not trustworthy, corresponding events are temporally close to each other. In other words, temporally close events have more probability to correspond to each other. Considering temporal correlation of the events, we have added internal dynamics into the cell activities such that each cell will preserve the last time it has been activated. Consequently, the contribution of each cell in the cooperative process can be weighted using a monotonically decreasing temporal kernel. From another point of view each cell keeps an internal dynamic by which its activity is fading over time like leaky neurons.
In consequence, the activity of each cell can be described by the following equations where W is temporal correlation kernel, E is the set of excitatory cells and I is the set of inhibitory cells for Cx, y, d^k, σ is a simple threshold function, α is a inhibition factor, and β tunes the slope of the temporal correlation kernel:

$$C^t_{x,y,d^k} = \sigma \left[\sum_{x'y'd' \in E} W^{t-t'}_{x'y'd'} C^{t'}_{x'y'd'} - \alpha \sum_{x'y'd' \in I} W^{t-t'}_{x'y'd'} C^{t'}_{x'y'd'} \right] \tag{5}$$

$$W^{\Delta t}_{x,y,d} = \frac{1}{1 + \beta \Delta t} \tag{6}$$

Hess [26] has analytically compared the *inverse linear* correlation kernel in Eq. (6) with *Gaussian* and *quadratic* kernels. He shows that this kernel can yield temporal correlation faster than Gaussian and quadratic functions without any obvious loss in quality.

Finally the disparity for a single event $e = (x_l, y_l, t_l)$ can be identified by Winner-Take-All mechanism over activity of the candidate cells:

$$D(e) = \arg_{d^k} \max \left\{ C^{t_l}_{x_l,y_l,d^k} \,\middle|\, C^{t_l}_{x_l,y_l,d^k} \ge \theta \right\} \tag{7}$$

General flow of the algorithm is depicted in Table 1. The following parameters shape the algorithm and need to be tuned:

- Excitatory neighborhood, r in Eq. (2): tunes the smoothness of the disparity maps.
- Inhibitory factor, α in Eq. (5): tunes the strength of inhibition during cooperation.
- Activation function threshold, θ: each cell is active if integrated input activity in Eq. (5) becomes larger than the threshold.
- Slope of temporal correlation kernel, β in Eq. (6): this parameter can adjust the temporal sensitivity of the cells to input events. Larger factor means faster dynamics and sharper temporal sensitivity to the upcoming events.

4 Experiments and Results

The experimental stereo setup that we use in this work is described in Chapter 2. The event packets are sent to a PC using two USB links, and the algorithm is implemented in MATLAB. There is no obvious standard benchmark in the literature to evaluate stereo matching algorithms using DVS. Rogister et al. used a moving pen as a simple stimulus which visually

Table 1 Flow of event-base cooperative stereo-matching algorithm

Algorithm.1, Event-driven cooperative stereo matching
Require: two synchronized retinas
Do for single event $e= (x, y, t)$
Construct set of possible corresponding candidates, $\mathbf{S_e}$ in equation (1).
for each corresponding pixel pair or equivalently $\{C_{x,y,dk} \mid 0 \leq d^k \leq d^{max}\}$
Find excitatory and inhibitory cells for $C_{x,y,dk}$, equations (2)-(4).
Compute activity of cooperative cell $C_{x,y,dk}$ according to equation (5).
End for
Do winner-take-all across all $C_{x,y,dk}$ cells.
If activity of winner cell is bigger than θ,
$D(e) = d^{WTA}$.
Else
Add small value ε to all corresponding cells activity.
End-If Update the cells, (time and activity).
End Do
Wait for next event.

showed the coherency of the detected disparity in depth [17]. To show the performance of the algorithm for temporally-overlapping stimuli, two simultaneously moving pens are used but the accuracy of the algorithm is not analytically reported [17]. Kogler et al. have used a rotating disk (similar to Fig. 1c) as a stimulus to analyze the detection rate in an area-based, an event image-based, and a time-based algorithm [16].

As our first experiment in this work, we create the disparity map of a single moving hand shaking in front of the retinas. In this experiment the algorithm has to deal with more complex stimuli than that of a single moving pen. The algorithm is executed without event buffering and the results for the stimulus located at 0.75 and 0.5 m are shown in Figs. 4 and 5 respectively. For better visualization in these figures, a stream of events is collected within 20 ms and two stereo frames are constructed in the left and the right retinas. Then the detected disparity for a single event is color-coded from blue to red for the disparity values varying from 0 to 40 pixels respectively. Moving the stimulus within two different known distances allows us to assess how coherent the detected disparity is with respect to the ground truth.

Since Rogister et al. have not quantitatively analyzed the performance of the algorithm in [17], we have replicated this algorithm. The parameters of this algorithm for each experiment are analytically set to achieve best results. Single-shot extracted disparity maps using two algorithms are shown in Figs. 4 and 5 (for the stimulus placed at 0.75 and 0.5 m respectively). As is depicted in the top row of Figs. 4a and 5a, the extracted disparity maps using the cooperative network are perfectly distinguishable, and as the objects come closer to the sensors, disparity is increased. Although the algorithm proposed in [17] is able to extract the disparity maps associated with the depths, the performance of this algorithm drops when the disparity is increased or equivalently stimuli come closer (compare Figs. 4a, 5a bottom). Moreover, the disparity map extracted using the cooperative network is sharper around the ground truth, as compared with the algorithm proposed by Rogister et al. [17]. In the top row of Figs. 4a and 5a, the coherency of the disparity maps with ground truth values is clearly depicted. The smoother maps extracted by the cooperative network are intrinsically provided by the local pattern of excitation in Eq. (2).

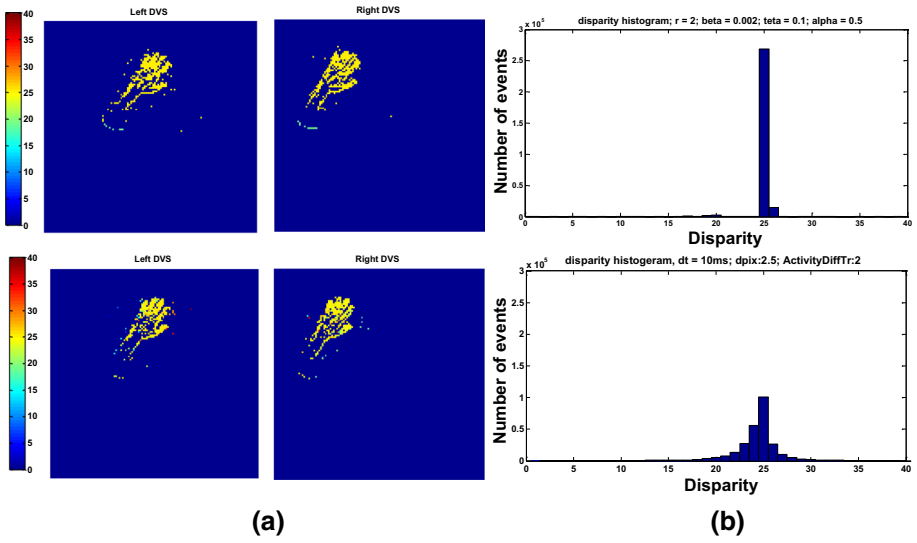

Fig. 4 **a** Color-coded disparity map of a 20 ms-long stream of events (in the left and the right retina) for a hand moving at 0.75 m, **b** detection histogram within time of 5 s. *Top row* the disparity maps and disparity histogram extracted by the cooperative network. *Bottom* extracted disparity maps and disparity histogram using algorithm in [17]

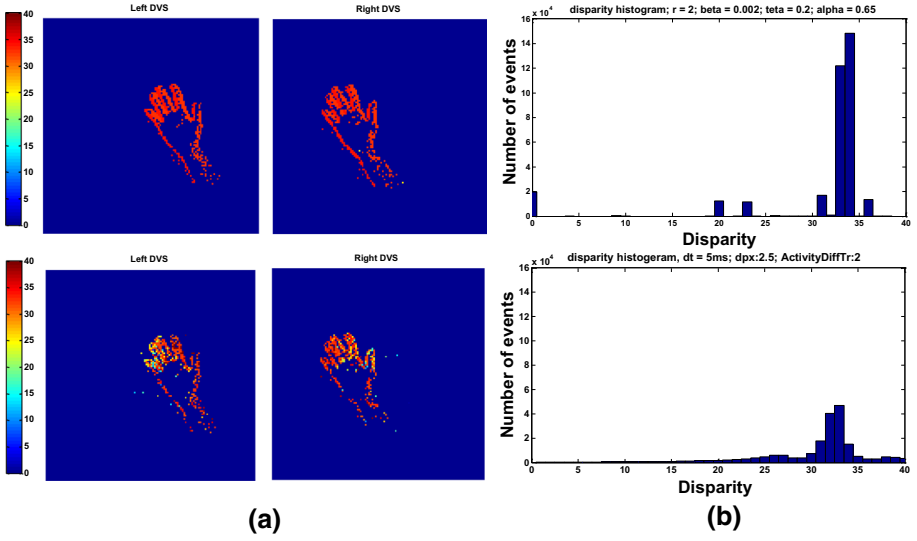

Fig. 5 **a** Color-coded disparity of a 20 ms-long stream of events (in the left and the right retina) for a hand moving at 0.5 m, **b** detection histogram within time of 5 s. *Top row* the disparity maps and disparity histogram extracted by the cooperative network. *Bottom* extracted disparity maps and disparity histogram using algorithm in [17]

Similar to the analysis performed in [16], and in order to analytically evaluate the detection rate, we have created the disparity histogram using both algorithms for the events generated within a time period of 5 s (Figs. 4b, 5b). The detection rate is the rate of the correct detected

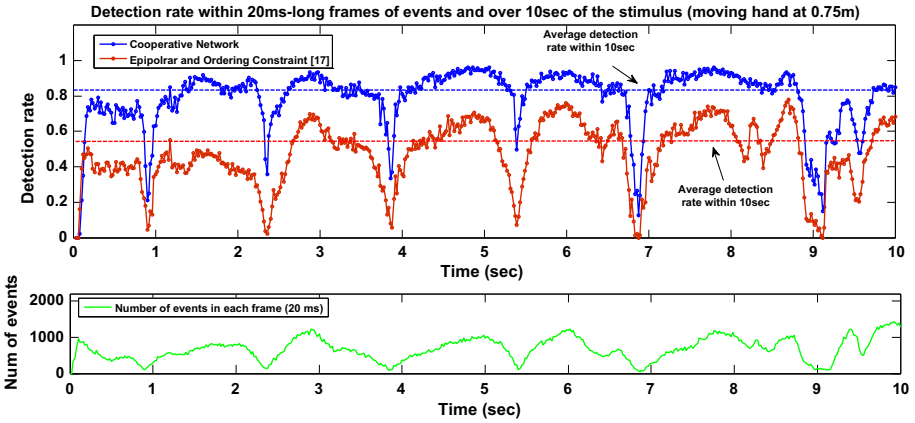

Fig. 6 *Top* Detection rate within 20 ms-long time bins (frames) and over 10 s of the stimulus (Moving hand at 0.75 m), *Bottom* number of events per time bin

disparity with respect to the ground truth and is used as a performance criteria in the previous works [16]. A range of detected disparity values within -1 and $+1$ of the ground truth value is considered as correct [16].

It is illustrated in Figs. 4b and 5b that, when the cooperative network is used, only a small fraction of the events are mismatched. For the moving hand at 0.75 m, 84 % of the events are perfectly mapped onto the disparity map 25 or 24 (Fig. 4b, top row), and for the stimulus located at 0.5 m, 74 % of the events are mapped onto the disparity maps 33 or 34 (Fig. 5b, top row). These results show the advantages of the cooperative approach compared to the purely time-based event matching, in which the best average detection rate for a simple stimulus does not exceed 30 % [16]. The average detection rates within 5 s and using the algorithm in [17] are 54 and 39 % respectively for the stimulus placed at 0.75 and 0.5 m (see the histograms at the Figs. 4b, 5b, bottom).

To analyze the detection rate over time, we have collected the stream of events detected within 20 ms-long time bins, and the detection rate for each time bin is calculated. The graphs of the detection rate within a time duration of 10 s for each experiment and using two algorithms are shown in Figs. 6 and 7. To compute detection rate in these graphs, the number of true matches divided by the number of whole events (including the events with unknown disparity) are calculated. For sparse time bins when the number of detected events has dropped, or equally when the stimulus is out of the overlapping retina's field of view, the momentary detection rate has dropped. This behavior is due the fact that when the stimulus is either partly located at the overlapping field of view, or it is out of the retina's field of view, a large number of events are detected as unknown disparity and the detection rate significantly decreases. But when the stimulus is located at both retina's field of views, the detection rate increases. The maximum detection rate of the algorithm proposed in [17] does not exceed 70 % for both experiments (red curve in Figs. 6 top, 7 top). Also it is clearly shown that the detection rate of the cooperative network is always higher as compared to previous work particularly in the sparse time bins (Fig. 7 top).

The results show that, the proposed network outperforms the algorithm proposed in [17], in which an exact event-by-event matching is performed and the epipolar and the ordering constraints are used in addition to temporal matching to enhance matching. For each single detected event, previous algorithms will search for a corresponding event in the other image,

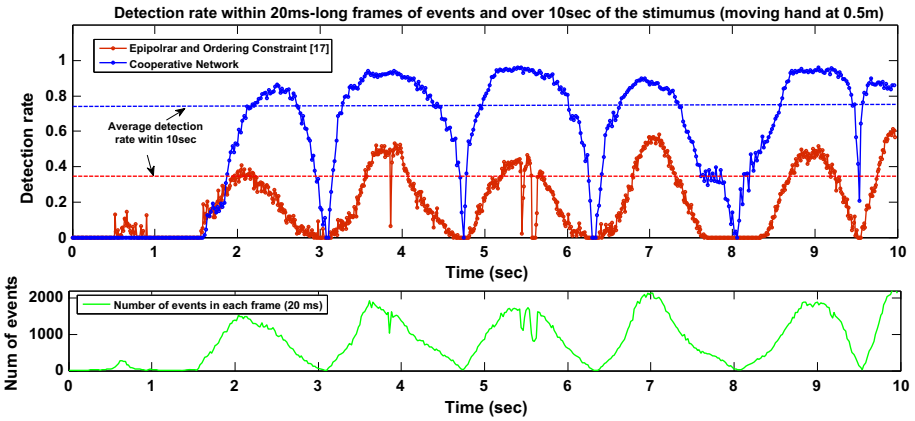

Fig. 7 *Top* Detection rate within 20 ms-long time bins (frames) and over 10 s of the stimulus (Moving hand at 0.5 m), *Bottom* number of events per time bin

whereas the proposed algorithm creates a distributed maps of the cells, through which the cooperative computation is performed over the most recent detected events. The activity of a single cell indicates the internal belief of the network in a specific matching pair. Each single event inserts a tiny piece of information into the network such that the belief in the false matches are suppressed. Enhanced detection rate of the cooperative network compared with previous works, is due to the computational power of the event-fusion matching versus exact event-by-event matching.

Comparing the detection histograms in Figs. 4b and 5b, the detection histograms for the stimulus located at 0.75 m is sharper around the ground truth as compared with the stimulus placed at 0.5 m. This shows there is more sensitivity of both algorithms to nearby objects and can be interpreted by the fact that in far distances, objects often generate few events. Thus, the correspondence problem should deal with less ambiguity for the objects moving in far distances and it is easier to find matching pairs. This behavior has been observed in previous works [16].

In order to evaluate the performance of the cooperative network in the cases with temporally-overlapping stimuli where the objects are located across common epipolar lines, we have created the disparity maps for two simultaneously moving hands, one is moving at 0.75 m and another one is moving at 0.5 m from the stereo DVS. In this scenario the algorithm should face considerably more ambiguity compared to the first experiment. The color-coded disparity values for a 20 ms-long stream of events, and the detection histogram within 5 s are presented in Fig. 8. As is depicted in this figure, the disparity maps which are purely computed in an event-based manner, is completely coherent with the depth of the moving objects (red color is corresponding to the events happened in 0.5 m and yellow shows the events detected at 0.75 m). In this experiment we have observed a considerable number of the events with unknown disparity (Fig. 8 bottom). Unknown disparity happens when the activity of the winner cell in Eq. (7) does not exceed the threshold θ. The increased rate of the unknown disparity in the second experiment is a result of the increased number of the events that are out of the overlapping field of view.

Previous works required additional constraints, e.g. ordering constraint, orientation, pixel hit-rate, and etc. to reduce the increased ambiguity of the temporally-overlapping events [17,19]. But in the cooperative network, the second pattern of inhibition [Eq. (4)] suppresses

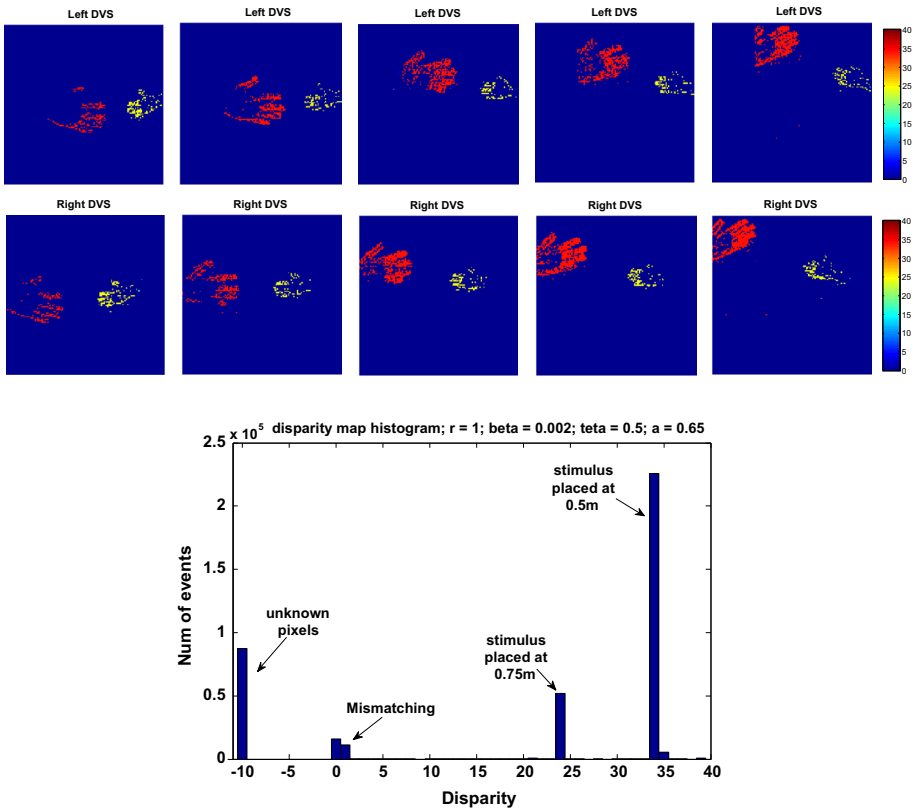

Fig. 8 *Top* color-coded extracted disparity maps over time for two moving hands (one at 0.75 m and another at 0.5 m). Each frame includes a stream of event generated within time of 20 ms. *Bottom* Detection histogram for stream of events generated in 5 s

a group of matching candidates that have been considered as corresponding pixels for a different object. This competitive process provides a mechanism to reduce false matches when multiple coincident clusters of events lie across or close to a common epipolar line and belong to different objects. In Fig. 9 the extracted disparity maps for two moving persons in a hallway is presented in time. In this experiment most of the events have the risk of multiple false matches, but the network can filter out a vast number of false events through the second pattern of inhibition.

The algorithm parameters for each experiment are listed in Table 2. Also it is worth mentioning that the algorithm is implemented in 64-bit *MATLAB 2012b,* running on an *Intel Core i5* 3.3 GHz processor with 16 GB RAM. The mean processing time per event with 25 % CPU load (one processor is fully loaded) is listed in Table 2 for each experiment. One important aspect in most stereo matching solutions is the processing time. The average processing time in the proposed algorithm particularly depends on the number of disparity maps d^{max}. If we observe a single event as an elemental feature, average required processing time for a network with 100 disparity maps does not exceed 1.5 ms per event on our computing platform. Although achieved processing time might be still acceptable for some applications, natural parallelism of the cooperative network can speed up the processing on parallel hardware, which can be addressed in future research.

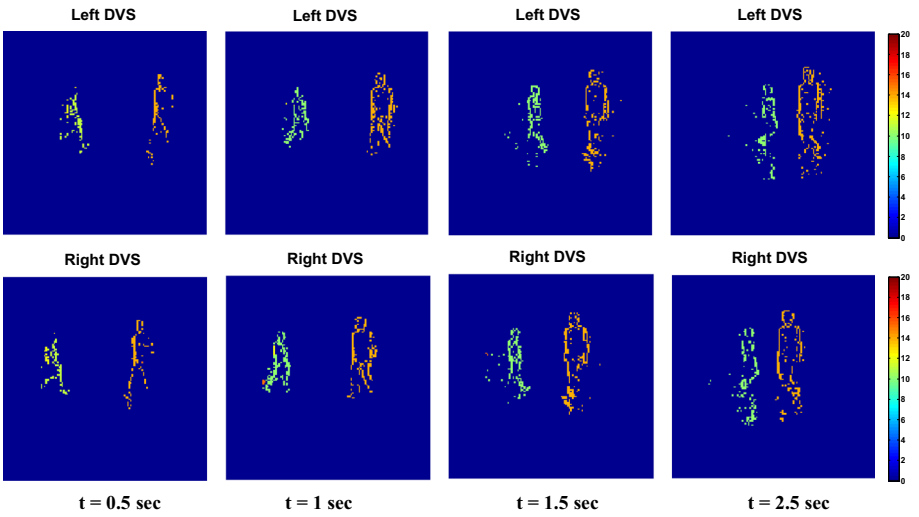

Fig. 9 color-coded extracted disparity maps over time period of 2.5 s for two persons moving in the hallway with different distances from the stereo DVS. The events are collected within a time history of 25 ms for each frame. *Green* disparity = 10; *Orange* disparity = 14. (Color figure online)

Table 2 Algorithm parameters for different experiments

Experiments/parameters	r	α	β	θ	d^{max}	Mean processing time per event (ms)
Single moving hand at 0.75 m	2	0.50	0.002	0.1	45	0.78
Single Moving hand at 0.5 m	2	0.65	0.002	0.2	45	0.81
Two moving hands (0.75 m & 0.5 m)	1	0.65	0.002	0.5	45	0.81
Moving in the hallway	1	0.65	0.0013	0.25	100	1.35

As a general rule of thumb, for sparse event-generating objects like the objects moving far away (or with slow motion), it is necessary to decrease the threshold of the cells activity θ (leads to more sensitivity of the cell), to allow sparse events to contribute in the cooperative process and not be cancelled as noise. Seemingly an adaptive homostacity mechanism [27] (i.e. adaptive θ) rather than global threshold setting, can help the network to detect sparse descriptive features. This work is worth to be investigated in future works.

5 Conclusion and Remarks

During the last decade, several attempts have been made to address the stereopsis problem using Neuromorphic Silicon Retina. Most of the existing stereo matching algorithms using DVS either are rooted in classical frame-based methods or temporal correlation. In order to fully take advantage of DVS sensors, developing efficient event-driven visual processing algorithms is necessary and remains an unsolved challenge. In this work we propose an asynchronous event-based stereo matching solution for DVS. The main idea of the proposed algorithm is grounded in cooperative computing principle which was first proposed by Marr in

the 80s. The classic cooperative approach for stereoscopic fusion operates on static features. The question we have addressed in this paper is how to formulate an event-driven cooperative process to deal with dynamic spatiotemporal descriptive features such as DVS events. To combine temporal correlation of the events with physical constraints, we have added a computationally simple internal dynamics into the network, such that each single cell can achieve temporal sensitivity to the events. Consequently the cooperation amongst distributed dynamic cells can facilitate a mechanism to extract a global spatiotemporal correlation for input events.

Knowing the disparity of a moving object in the retina's field of view, we have used two basic experiments to analyze the accuracy and the detection rate of the proposed algorithm. Obtained disparity maps are smooth and coherent with the depth in which the stimuli are moving. The detection rate considerably outperforms previous works. In the second experiments we evaluated the performance of the algorithm in response to temporally-overlapping events. The results show that the cooperative dynamics intrinsically reduces the ambiguity of the correspondence problem when coincident cluster of events lie on the same epipolar lines.

Acknowledgments The authors would like to cordially thank *Christoph Richter* and *Marcello Mulas* for their helpful comments on this article, and *Nicolai Waniek* for his fruitful discussions on vision science.

References

1. Bruce V, Green PR, Mark A (2003) Visual perception: physiology, psychology, & ecology. Psychology Press, New York
2. Indiveri G, Douglas R (2000) Neuromorphic vision sensors. Science 288(5469):1189–1190
3. Wen B, Boahen K (2009) A silicon cochlea with active coupling. IEEE Trans Biomed Circuits Syst 3(3):444–455
4. Furber SB, Brown AD (2009) Biologically-inspired massively-parallel architectures—computing beyond a million processors. In: 9th international conference on application of concurrency to system design (ACSD), USA, 1–3 July 2009, pp 3–12
5. Marr D (2010) Vision, a computational investigation into the human representation and processing of visual information. In: Vaina Lucia M (ed) MIT press, Cambridge. http://mitpress.mit.edu/books/vision-0
6. Mahowald M, Delbrück T (1989) Cooperative stereo matching using static and dynamic image features. In: Mead C, Ismail M (eds) Analog VLSI implementation of neural systems. Kluwer Academic Publishers, Boston, pp 213–238
7. Hartly R, Zisserman A (2003) Multiple view geometry in computer vision, 2d edn. Cambridge University Press, Cambridge
8. Conradt J, Pescatore M, Pascal S, Verschure PFMJ (2002) Saliency maps operating on stereo images detect landmarks and their distance. In: International conference on artificial neural networks (ICANN), Madrid, Spain, 2002, pp 795–800
9. Ventroux N, Schmit R, Pasquet F, Viel PE, Guyetant S (2009) Stereovision-based 3D obstacle detection for automotive safety driving assistance. In: 12th international IEEE conference on intelligent transportation systems, Oct 2009, pp 1–6
10. Lichtsteiner P, Posch C, Delbruck TA (2008) A 128 $\times$ 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE J Solid State Circuits 43(2):566–576
11. Conradt J, Cook M, Berner R, Lichtsteiner P, Douglas RJ, Delbruck T (2009) A pencil balancing robot using a pair of AER dynamic vision sensors. In: IEEE international symposium on circuits and systems, Taipeh, Taiwan, May 2009, pp 781–784

12. Drazen D, Lichtsteiner P, Hafliger P, Delbruck T, Jensen A (2011) Toward real-time particle tracking using an event-based dynamic vision sensor. Exp Fluids 51:1465–1469
13. Müller GR, Conradt J (2012) Self-calibrating marker tracking in 3D with event-based vision sensors. In: 22nd international conference on artificial neural networks (ICANN), pp 313–321
14. Ni Z, Pacoret C, Benosman R, Ieng S-H, Régnier S (2012) Asynchronous event-based high speed vision for microparticle tracking. J Microsc 245(3):236–244
15. Waniek N, Bremer S, Conradt J (2014) Real-time anomaly detection with a growing neural gas. In: 26th international conference on artificial neural networks (ICANN), Hamburg, Germany, pp 97–104
16. Kogler J, Humenberger M, Sulzbachner C (2011) Event-based stereo matching approaches for frameless address event stereo data. In: Advances in visual computing, Lecture notes in computer science, vol 6938, pp 674–685
17. Rogister P, Benosman R, Ieng SH, Lichtsteiner P, Delbruck T (2012) Asynchronous event-based binocular stereo matching. IEEE Trans Neural Netw Learn Syst 23(2):347–353
18. Carneiro J, Ieng S, Posch C, Benosman R (2013) Event-based 3D reconstruction from neuromorphic retina. Neural Netw 45:27–38
19. Camuñas-Mesa LA, Serrano-Gotarredona T, Ieng SH, Benosman R, Linares-Barranco B (2014) On the use of orientation filters for 3D reconstruction in event-driven stereo vision. Front Neurosci 8:48
20. Mahowald MA, Mead C (1991) The silicon retina. Sci Am 264(5):76–82
21. Hoffmann R, Weikersdorfer D, Conradt J (2013) Autonomous indoor exploration with an event-based visual SLAM system. In: European conference on mobile robots, Barcelona, Spain, Sep 2013, pp 38–43
22. Weikersdorfer D, Adrian DB, Cremers D, Conradt J (2014) Event-based 3D SLAM with a depth-augmented dynamic vision sensor. In: IEEE international conference on robotic and automation, Hong Kong, June 2014, pp 359–364
23. Galluppi F, Denk C, Meiner MC, Stewart T, Plana L, Eliasmith C, Furber C, Conradt J (2014) Event-based neural computing on an autonomous mobile platform. In: IEEE international conference on robotics and automation (ICRA), Hong Kong, China, 31 May-7 June 2014, pp 2862–2867
24. Conradt J, Galluppi F, Stewart TC (2015) Trainable sensorimotor mapping in a neuromorphic robot. Robot Auton Syst. doi:10.1016/j.robot.2014.11.004
25. Zhu Y, Quian N (1996) Binocular receptive fields, disparity tuning and characteristic disparity. Neural Comput 8:1647–1677
26. Hess S (2006) Low-level stereo matching using event-based Silicon Retina. Semester project report, ETHZ Zurich, p 16
27. Remme MW, Wadman WJ (2012) Homeostatic scaling of excitability in recurrent neural networks. PLoS Comput Biol 8(5):e1002494