2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 978-1-6654-4899-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/CVPRW53098.2021.00154

A Cortically-inspired Architecture for Event-based Visual Motion Processing: From Design Principles to Real-world Applications

Francesca Peveri
francesca.peveri@edu.unige.it

Simone Testa
simone.testa@edu.unige.it

Silvio P. Sabatini
silvio.sabatini@unige.it

Department of Informatics, Bioengineering, Robotics and Systems Engineering
University of Genoa - via Opera Pia 11a, Genoa, ITALY

Abstract

We developed and tested the architecture of a bio-inspired Spiking Neural Network for motion estimation. The computation performed by the retina is emulated by the neuromorphic event-based image sensor DAVIS346 which constitutes the input of our network. We obtained neurons highly tuned to spatial frequency and orientation of the stimulus through a combination of feed-forward excitatory connections modeled as an elongated Gaussian kernel and recurrent inhibitory connections from two clusters of neurons within the same cortical layers. Sums over adjacent nodes weighted by time-variable synapses are used to attain Gabor-like spatio-temporal V1 receptive fields with selectivity to the stimulus' motion. In order to gain the invariance to the stimulus phase, the two polarities of the events provided by the neuromorphic sensor were exploited, which allowed us to build two pairs of quadrature filters from which we obtain Motion Energy detectors as described in [2]. Finally, a decoding stage allows us to compute optic flow from the Motion Detector layers. We tested the approach proposed with both synthetic and natural stimuli.

1. Introduction

We are immersed in a world in constant motion. For this reason visual motion perception has been the subject of extensive research in the fields of perceptual psychology, neuro-physiology, and computational vision. Widespread evidence pointed out that the mammalian brain evolved specific mechanisms responsible for the processing of visual stimuli in order to extract motion information [8, 16]. These mechanisms characterize the cortical "motion pathway" that involves first and foremost, the primary (or striate) visual cortex (area V1), and, then, continues to the middle temporal visual area (MT or V5) and other extrastriate areas. According to the principle of "building-to-comprehend", advances in high-performance computing is the best ally to understand the properties of cells belonging to these cortical areas. Recent asynchronous event-driven cameras combined with Spiking Neural Networks (SNNs) allow, indeed, real time simulations of large scale neural networks by monitoring and manipulating any variable in each neuron or synapse.
From an application perspective, asynchronous event-based artificial retinas and SNN processors turn out to be the best solutions to achieve flexibility and real-time performance. Dynamic vision sensors [13] do not provide conventional frames, but asynchronous ON and OFF events that signal scene reflectance changes. These continuous-time sensors functionally emulate some key features of the human retina and represent a major shift from conventional frame-based sensors, owning to the advantages of high temporal resolution and low power consumption. Accordingly, they transmit only pixel-level changes, at microsecond time, equivalent to a high-speed camera at thousands of frames per second, but with far less data.

The price we pay when following this approach lies in the necessity of specific novel algorithmic solutions, since those developed for frame-based visual input are no longer applicable. Plausible design solutions can be inspired by the computational paradigms adopted in the visual cortices, as they conform more to intrinsic structural properties of the visual signal than to abstract calculus (i.e., to the computational theory of the problem). In general, early vision perceptual processes can be interpreted as a "measuring" operation on a visual signal [1], by which to extract, on a local basis, specific characteristics of the signal (amplitude and phase of spectral components, orientation, direction of motion, etc.). In such a way, a parametric representation of input signal occurs, on which to base the sub-

978-1-6654-4899-4/21/$31.00 ©2021 IEEE
DOI 10.1109/CVPRW53098.2021.00154

sequent interpretation of the events of the visual scene, by eventually combining such local representation over larger spatial neighborhoods. In this scenario, guided by neurophysiological and modeling findings on the properties of visual cortical neurons, a first challenge is to demonstrate how a variety of computational visual tasks can be built by following a compositional approach, which hierarchically provide more complex visual descriptors by combining a limited set of stereotyped basic blocks. Interestingly, these solutions can have direct correspondences with their classical counterparts, and several works (e.g., [15, 10]) have demonstrated that firing rate models can properly solve the computational problem with adequate efficacy, higher flexibility, and robustness to adverse illumination conditions and to low S/N ratios. Yet, a second and harder challenge is to demonstrate that the design principles exploited in the large-scale network of firing-rate model neurons can be equivalently mapped through an event-based coding on SNNs.

In this paper, we faced this challenge by designing and testing a complete multilayer SNN for an explicit motion estimation, i.e., the optic flow, in order to assess the network performances through realistic motion sequences. The network functionally mimics the cortical motion pathway. The proposed neural architecture refers to a variant of the Heeger and Simoncelli model [12][23] and is described through a three-layer architecture composed of distributed populations of cells. In the first layer, Gabor-like spatial receptive fields allow a band-pass filtering on the events coming from a neuromorphic sensor, the DAVIS346, in order to extract early vision features as the spatial frequency and orientation from a dynamic stimulus. In the second layer, a bank of spatio-temporal oriented filters approximates the receptive fields (RFs) of the simple cells' population of area V1, which are tuned to different motion directions and contribute to build the population of complex cells as motion energy units [2]. The responses of the complex cells are combined in the third layer to obtain estimates of the magnitude and direction of local velocities, as it happens in area MT [21]. Unlike other neuromorphic solutions that are far from a bio-inspired approach, our network tries to faithfully emulate cortical processing, both in the computational paradigms and in the dynamics of individual neurons, which can be used as *building blocks* to design visual feature detectors of increasing complexity.

2. Related work

Our work is positioned within the framework of bio-inspired neuromorphic systems, in which the use of event-driven sensors is making its way. Nowadays, the high demand of these sensors is due to the great advantage offered in real-time computing, especially for robotics, self-driving vehicles and wearable systems. Many other works have proposed SNNs capable of estimating speed and direction of motion and the extraction of the optic flow. The first approaches attempted to adapt well-known solutions in the field of computer vision to an event-based framework. For example, Benosman *et al.* [6] translated, in an event-based approach, the constraints from one of the most popular techniques formulated by Lucas and Kanade [14], based on the brightness constancy assumption. In [9] the authors proposed a hierarchical architecture for optical flow estimation that use a bank of spatio-temporal filters selective at different speeds and directions of motion (Gabor filters), equivalent to correlation detectors. A more bio-inspired approach is described in [17], in which an unsupervised SNN implements a novel spike timing-dependent plasticity (STDP) rule, in order to learn the proper filters from event data. Barranco et al. [4] proposed a simple method for locating texture regions and a novel phase-based method for motion estimation. A remarkable work is [20], in which the authors compared the accuracy and processing time of nine event-based optical flow algorithms. The algorithms considered were a direction selective filter [11], four variants of the Lucas-Kanade algorithm, four variants of local plane fits [5], and a flow estimation based on the camera's gyro information instead of visual motion cues. In order to comparatively evaluate these methods, the authors created a public dataset composed of synthesized samples and real samples recorded from a 240x180 pixel Dynamic and Active-pixel Vision Sensor (DAVIS). Finally, mention goes to [25], in which cortical mechanisms combining filters with spatio-temporal tuning are emulated and also used for classification purposes.

Undoubtedly, the greatest challenge in this specific area is to propose solutions that are biologically plausible, in order to have the double advantage of reflecting computations actually present in the cortex and to better understand their functioning through emulation. Clearly, they should also have the proper features for an efficient implementation in neuromorphic processors.

3. Cortical-style visual processing

The proposed neural architecture refers to a variant of the Heeger and Simoncelli model [12][23] and is described through a three-layer network composed of distributed populations of cells. In the first layer, Gabor-like spatial receptive fields (RFs) allows a band-pass filtering on the input. In the second layer, a bank of spatio-temporal oriented filters approximates the RFs of simple cells in area V1. These neurons are tuned to different motion directions and contribute to build the population of complex cells as motion energy units [2]. The responses of the complex cells are combined in the third layer to obtain estimates of the magnitude and direction of local velocities, as in MT cortical area [21].

The schematic representation in Fig. 1 shows the entire ar-

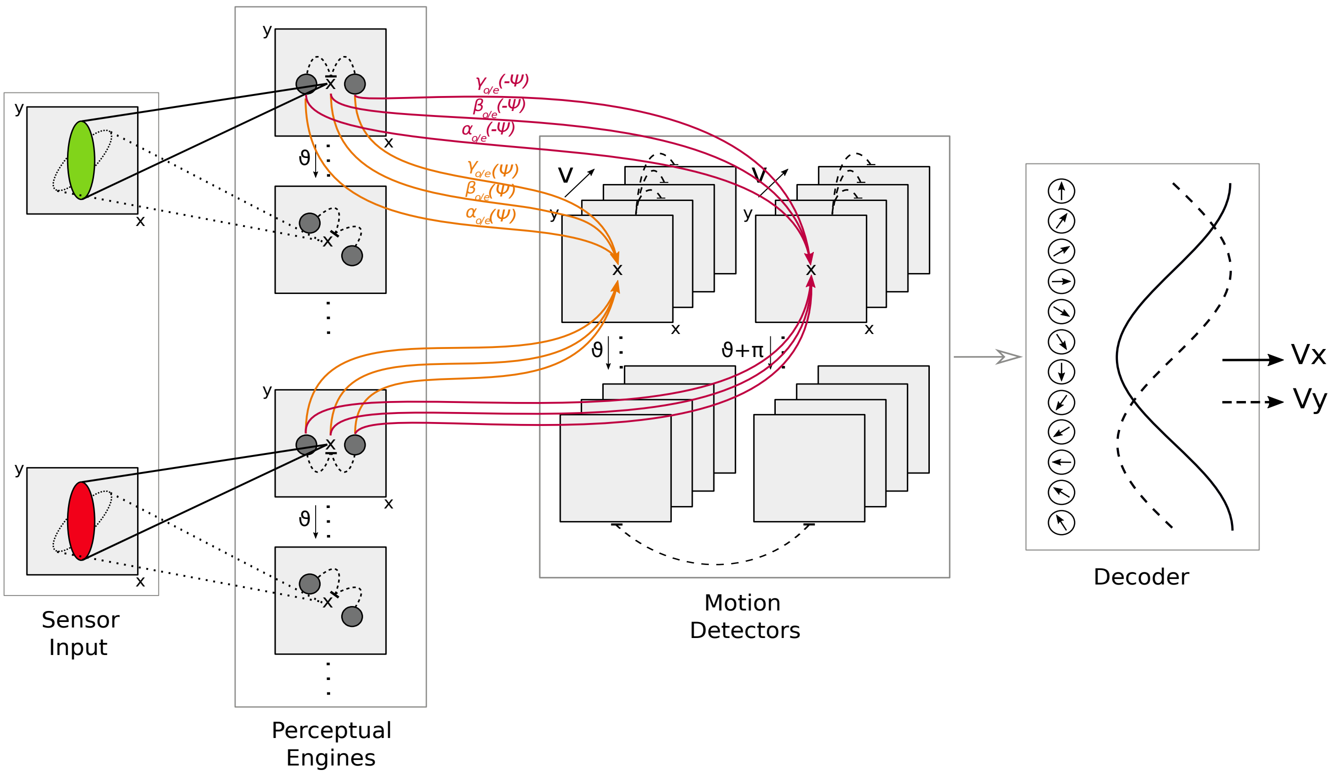

Figure 1. **Overall network architecture.** The network is composed of (i) an input stage where the events, provided by the sensor, are split in the two polarities (ON and OFF), (ii) the Perceptual Engines, (iii) a Motion Detector stage and (iv) a decoding stage for velocity components. All synaptic connections (both solid and dashed lines) starting from the *sensor input* and ending in the *perceptual engines* are excitatory. The dotted lines in the latter, indicating the recursive connections, are inhibitory. In the third layer of *motion detectors*, the dotted lines indicate inhibitory synaptic connections.

chitecture of the network. It is composed of three main building blocks: (i) an input stage, (ii) perceptual engines and (ii) motion detector units. The final block represent the decoding process.

Retinal input Starting from the sensory input, events in the pixel array are divided in two groups according to their ON or OFF polarity. This distinction will serve to create, in subsequent stages, two pairs of quadrature filters in order to obtain invariance on the stimulus' phase, an essential component of the Energy Model [2]. Squaring the response of quadrature cells acting on a gray-scale image is no longer feasible in a spike-based encoding mechanism. Therefore, pairs of counterphase receptive fields are used on both ON and OFF events, which separately encode the increase and decrease of the light stimuli, respectively.

Oriented perceptual engines In the second stage, the two parallel ON and OFF channels are preserved. Both channels share the same configuration: same feed-forward convolutional kernels acting on input events and same clustered recursive inhibitions. Specifically, the weights of the excitatory feed-forward synapses are arranged and shaped according to an elongated Gaussian function. This provides the neurons of the second stage with an orientation preference. Here, neurons in both channels also receive recurrent inhibitory afferences from two laterally clustered groups of cells of the same layer. These groups are located at a distance **d** along the direction orthogonal to the major axis of the feed-forward kernel. This recurrent inhibitory schema, better described in [22], gave us the great advantage of reducing the number of required connections to obtain neurons with highly-structured receptive fields that approximate two-dimensional Gabor functions, markedly tuned to stimulus' spatial frequency and orientation. Neurons' sensitivity can be controlled and sharpened by appropriately changing the spatial distance and the extensions of these inhibitory kernels.

The neurons in this second stage are hereafter referred to as “perceptual engines” because they provide computational primitives that can be composed to obtain more powerful feature extractors by simply adjusting the weights of specific efferences. The excitation of a neuron at this stage, with orientation θ and in position $\mathbf{n} = (n_x, n_y)$, could be formulated as:

$$e(\mathbf{n}) = a \sum_{p \in R} h_{ff}(\mathbf{n}-\mathbf{p}) s(\mathbf{p}) - b \sum_{q \in C} h_{fb}(\mathbf{n}-\mathbf{q}) e(\mathbf{q}) \quad (1)$$

where a and b are respectively the strength of the feedforward (h_{ff}) and recurrent (h_{fb}) kernels, s is the visual stimulus, R is the domain of the retinal driving input, and C is the domain of the lateral cortical inhibition. Such an equation describes the input/output relation of a linear recurrent inhibitory network characterised by a direct (i.e., forward) feeding from an oriented neighborhood on the retina layer and by clustered recurrent inhibitory contributions from neurons lying along the orthogonal direction. The resulting spatial impulse response h of neuron $\mathbf{n}$ is characterised by an even-symmetric spatial profile. The final neuron 'firing rate' response is obtained through a rectified linear activation function: $r(\mathbf{n}) = F[e(\mathbf{n})]$.

We can consider the function $h(\mathbf{n})$ as a primary characteristics of the network and it represents an eigenmode (the *perceptual engine*) that can be used to obtain more powerful network responses. It has been demonstrated [19] that by combining the responses of three neurons of the V1 layer, the central one in position $\mathbf{n}$ and the lateral ones at positions $\mathbf{n} - \mathbf{d}$ and $\mathbf{n} + \mathbf{d}$, it is possible to obtain Gabor-like functions of any phase ψ. The resulting filter therefore represents a good approximation of a typical kernel useful for computational vision processing:

$$\begin{aligned} g(\mathbf{n}) = \alpha h(\mathbf{n}-\mathbf{d}) + \beta h(\mathbf{n}) + \gamma h(\mathbf{n}+\mathbf{d}) = \\ \cong C e^{-\mathbf{n}\cdot\mathbf{n}/\sigma^2} cos(\mathbf{k_0} \cdot \mathbf{n} + \psi) \end{aligned} \quad (2)$$

where $\mathbf{k_0}$ represents the preferred spatial-frequency vector of the neuron (with direction given by the tuning orientation θ) and σ the extension of its receptive field. The phase ψ can be then adjusted as desired by simply setting appropriate values to the weights α, β and γ. A general expression for these variables, as functions of the phase ψ, can be formalized as follows:

$$\begin{cases} \alpha = -Bsin(\psi) - Acos(\psi) \\ \beta = cos(\psi) \\ \gamma = Bsin(\psi) - Acos(\psi) \end{cases} \quad (3)$$

where A and B are coefficients whose values can be chosen as 0.5 and 1, respectively. In this way, we ensure that the sum of the three weights is equal to zero, i.e., at least ideally, the resulting RF will filter out the DC component.

For any given value of the spatial phase ψ, we will then have three specific values for the weights. As an example, consider the case of orthogonal responses (i.e. a quadrature pair) $\mathbf{g}(\mathbf{n}) = g_c(\mathbf{n}) + jg_s(\mathbf{n})$. The two symmetries can be straightforwardly obtained by posing $\alpha = -0.5$, $\beta = 1$, $\gamma = -0.5$ for the even response, while $\alpha = -1$, $\beta = 0$, $\gamma = 1$ for the odd one.

Motion detectors By introducing a temporal dependency in the weighting coefficients α, β, and γ of equation (3), it is possible to obtain spatio-temporal filter. This mechanism leads to a Gabor-like kernel with a time-variable spatial phase $\psi(t) = \omega_0 t$, as a local travelling wave with constant velocity $\mathbf{v}_f$. Such a filter will therefore be able to detect the presence or absence of a stimulus moving along a specific direction and at a specific speed.

The resulting filter is described by the following equation:

$$g(\mathbf{n}, t) = C' e^{-t/\tau} e^{-\mathbf{n}\cdot\mathbf{n}/\sigma^2} cos(\mathbf{k_0} \cdot \mathbf{n} \pm \omega_0 t) \quad (4)$$

with $\omega_0 = \mathbf{v}_f \cdot \mathbf{k_0}$, where w_0 and $\mathbf{k_0}$ are the preferred temporal and spatial frequencies respectively and $\mathbf{v}_f$ is the tuning velocity of the neuron; the preferred direction of motion along θ is determined by the sign of $\psi(t)$. Finally, the term $e^{-t/\tau}$ represents a temporal envelope defined by the synaptic integration with time constant τ.

A quadrature pair of spatio-temporal filters can therefore be directly obtained by considering two sets of synaptic connections, for even and odd symmetries respectively, that weight the responses of the three mentioned above neurons. In Fig. 1 the subscripts e/o of the weight functions refer the different symmetries. Specifically, in order to obtain an odd symmetry, the argument of the cosine function in equation (4) should have an additional phase shift of $-\frac{\pi}{2}$. Overall, a pair of quadrature filters was designed from both ON and OFF channels, to achieve phase-invariant responses in the post-synaptic neurons. Obviously, the addition of further synapses with weights leading to different phase shifts would improve the invariance property of such neurons to the phase of the stimulus.

Due to these peculiar combinations of perceptual engines, we obtain the “Motion Detector” units:

$$\begin{aligned} E(\mathbf{n}, t; \theta, v_f) &= r_c^{\text{ON}}(\mathbf{n}, t; \theta, v_f) + r_s^{\text{ON}}(\mathbf{n}, t; \theta, v_f) \quad (5) \\ &+ r_c^{\text{OFF}}(\mathbf{n}, t; \theta, v_f) + r_s^{\text{OFF}}(\mathbf{n}, t; \theta, v_f), \end{aligned}$$

which constitute the third stage of our architecture. To refine the tuning properties on both the velocity magnitude and direction of motion, a competition mechanism has been introduced via soft winner-takes-all (WTA). In more detail, each neuron inhibits all other neurons having different velocity (or motion-direction) selectivity but with RF centered on the same retinal location.

Decoding strategy From the *motion detectors* stage it is possible to decode the motion energy responses along each spatial orientation to compute component velocities v_θ. In order to estimate the component velocity v_θ along the preferred orientation of the cells, we use a center of mass approach described in [18]:

$$v_\theta(\mathbf{n}, t; \theta) = \frac{\sum_i^N v_{fi} G(\mathbf{n}) * E(\mathbf{n}, t; \theta, v_{fi})}{\epsilon + \sum_i^N G(\mathbf{n}) * E(\mathbf{n}, t; \theta, v_{fi})} \tag{6}$$

where $E(\mathbf{n}, t; \theta, v_{fi})$ is the output of the motion energy detector tuned to the speed v_{fi} (i.e., in a spike-based framework, the instantaneous firing rate of such neuron), $G(\mathbf{n})$ is a Gaussian window used for pooling the motion detector over a spatial neighborhood, and ϵ is a positive (small) constant that prevents division by zero.

Due to the aperture problem [3, 7], a motion estimation based on the local computation of oriented RFs can recover only the velocity component that is perpendicular to the filter orientation. To this purpose, a measure of the full velocity vector $\mathbf{v_p} = (v_x, v_y)$ is achieved by means of an intersection-of-constraints (IOC) mechanism. The individual components of such vector are estimated as described in [10], according to which the least square solution of the IOC-based formulation for the computation of full velocity can be written as:

$$\begin{aligned} v_x(\mathbf{n}, t) &= \frac{2}{N} \sum_{\theta_i=\theta_1}^{\theta_N} v_{\theta_i}(\mathbf{n}, t) cos(\theta_i) \\ v_y(\mathbf{n}, t) &= \frac{2}{N} \sum_{\theta_i=\theta_1}^{\theta_N} v_{\theta_i}(\mathbf{n}, t) sin(\theta_i) \end{aligned} \tag{7}$$

The resulting velocity field, together with its confidence - represented by the instantaneous firing rates of the spiking motion detectors - represents the estimated optic flow.

4. Experiment

4.1. Event-based dataset

For characterizing the behavior of single neurons, we used a set of stimuli consisting of drifting gratings with different orientations θ and speeds v_s. A drifting grating is a sinusoidal oscillation in luminance L that moves at a constant velocity (whose magnitude is determined by ω_s) along the direction of the wave vector $\mathbf{k}_s$:

$$L(\mathbf{x}) = m[1 + c\ sin(\mathbf{k_s} \cdot \mathbf{x} + \omega_s t)] \tag{8}$$

where $\mathbf{x}$ denotes the spatial domain, m the mean background luminance and $c \in [0, 1]$ the spatial contrast. This type of stimulation allows us to study the response of a neuron to variations of $\mathbf{v_s} = \omega_\mathbf{s} \mathbf{k_s}$ and $\mathbf{k_s}$ (both modulus and direction θ).

The moving grating was presented on a computer monitor with resolution 1920 × 1080, refresh rate of about 144 Hz, and maximum brightness. The DAVIS346 event-based camera, used for the recordings, was placed in front of the monitor at a distance of 30 cm, in a specifically-dedicated dimly lit room. The Python code handling the simultaneous presentation of the moving stimulus, together with the recording of the output events from the sensor, leverages the multiprocessing technique. The communication with the neuromorphic camera is based on a serial connection and all events were saved to disk as numpy arrays for off-line processing. The computer managing both data logging and stimuli display run under Ubuntu-Linux 20.04 operating system. All recordings lasted 2 seconds and were performed by setting the neuromorphic sensor biases to the default values. The pixel array was then cropped by taking only the central 100 × 100 portion, in order to limit the computational cost of the subsequent network simulation.

In all the experiments the gratings' contrast c was kept constant to the maximum value. Stimuli were presented with 24 orientations evenly spaced in range $[0 - 180)$ deg, with a 15 deg step. The spatial frequency values ranged from 0 to 1.6 cyc/deg with a 0.2 cyc/deg constant step. Finally, the following values were chosen for the stimulus velocities: 1, 2, 3, 4 deg/sec.

4.2. Simulations

The network was simulated for 2 seconds, with a simulation time-step of 0.1 ms. In order to characterize the motion detectors, we consider all orientations and speeds of the grating stimuli for the fixed spatial frequency value of 0.6 cyc/deg. Instead, for characterizing the perceptual engines alone, we took all spatial frequencies.

In the context of this work, the simulator used was *Brian2* [24], an open source, intuitive and highly flexible tool for spiking neural networks. The neuron model we chose to adopt is an Adaptive Exponential Integrate-and-Fire neuron model (AdEx). The AdEx model can produce many complex firing patterns observed in biology by tuning a limited number of parameters, e.g. spike-frequency-adaptation, bursting, regular/irregular spiking and transient spiking. The evolution of the membrane potential in the AdEx model is described by a two-variable equation as below:

$$C_m \frac{dV_m}{dt} = -g_L(V_m - E_L) + g_L \Delta_T e^{\frac{(V_m - V_T)}{\Delta_T}} - w + I \tag{9}$$

where V_m is the membrane potential, I is the input (post-synaptic) current, C_m the membrane capacitance, g_L the leak conductance, E_L the leak reversal potential, V_T the

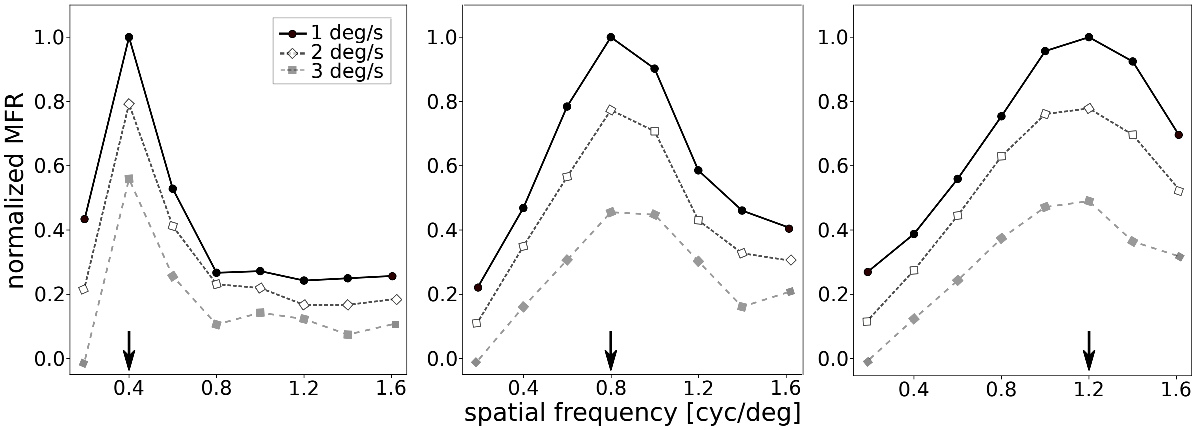

Figure 2. **Spatial frequency tuning in perceptual engines.** Three tuning curves obtained from the perceptual engines stage. The black arrows point to the spatial frequency value at which the population is most selective. The abscissas show the spatial frequency values and the ordinates the normalized mean firing rate of the population.

threshold and Δ_T is the slope factor. The adaptation current w has its own temporal dynamics described by the equation:

$$\tau_w \frac{dw}{dt} = \eta(V_m - E_L) - w \tag{10}$$

where η is the adaptation coupling parameter and τ_w is the adaptation time constant. If the membrane voltage crosses a certain threshold voltage V_T, a spike is emitted and the neuron is reset:

$$\begin{aligned} V &\rightarrow V_{reset} \\ w &\rightarrow w + \kappa \end{aligned} \tag{11}$$

where the parameter κ is responsible for spike-triggered adaptation. Note that the AdEx model can be simply reduced to the standard Leaky Integrate-and-Fire model by taking the limit $\Delta_T \rightarrow 0$ in equation (9) and deactivating the coupling parameter in (10).

Concerning the synaptic transmission, the biological mechanism follows a complex but well established process. It is challenging to model such a process and its dynamics. Nevertheless, many simple phenomenological models of synapses can represent the time and voltage dependence of synaptic currents fairly well. Therefore, we opt for an *exponential function* in order to model the synaptic dynamics. This functions describes the evolution of the synaptic conductance and hence the dependence of the post-synaptic current to an input spike at time t_0:

$$g_{syn}(t) = \bar{g}_{syn} exp\left(-\frac{t - t_0}{\tau}\right) \tag{12}$$

Because of the presence of a single time constant τ, the rising phase is instantaneous while the decay phase follow the exponential term. In general this is far from a biological condition, however provides a reasonable approximation for many synapses.

5. Results

Spatial frequency tuning Firstly, we tested the perceptual engine stage of the network with a set of drifting gratings having different spatial and temporal frequencies. The aim was to ensure that the population, through the combination of feed-forward and recurrent contributions, was capable of acquiring selectivity at a specific spatial frequency for all possible speeds of the input stimuli. As predicted by the firing-rate model [22], the parameters that mostly influence such tuning are the geometrical dimensions of both feed-forward and recurrent Gaussian kernels. Particularly, the major impact is given by the standard deviation of both kernels and the distance $\mathbf{d}$ between the two inhibitory clusters. The resulting tuning curves, obtained by averaging the activity of the entire cells' population in the perceptual engine stage, are shown in Fig. 2. Changing the relevant parameters, we can obtain different spatial frequency selectivity. We report three tuning examples at 0.4 cyc/deg, 0.8 cyc/deg and 1.2 cyc/deg. This behaviour is reproducible for all tested speeds of the gratings.

Motion selectivity In order to make neurons selective to the motion direction of the stimulus and to the desired set of speeds $\mathbf{v}_f$, we had to impose the proper temporal frequencies w_0 to the time-variable synapses by acting on the synaptic weights of equation (3). Since the tuning spatial frequency of the neurons was set to 0.6 cyc/deg, in order to achieve preferences to the speeds ± 1, 2, 3 and 4 deg/s, we had to specify the following temporal frequencies: ± 0.6, 1.2, 1.8 and 2.4 cyc/s. Figure 3 shows the activity of motion detectors with 4 different tuning speeds relative to all possible speeds of the presented stimuli. The left plot shows the selectivity for positive speeds while the right one for the negative motion. We can therefore notice that, as in the classical energy model [2], our motion detectors are able to discriminate fairly well the direction in which the stimulus is moving and (although with some variability) the speed value. Note that no decoding strategy was adopted to plot these curves, but we only took the average activity of sets of motion detector neurons.

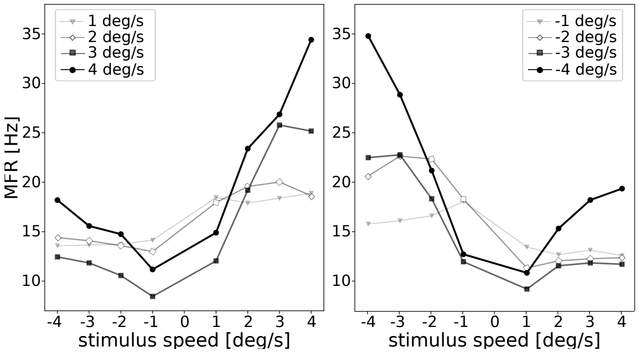

Figure 3. **Speed tuning in motion detectors.** The network was fed with events relative to drifting gratings with both positive (on the left) and negative speeds (on the right) at different values. The peak of each curve indicates the speed at which the population is more selective. Colour opacity decreases as the modulus of the speed decreases

Since we have considered different possible tuning orientations for the neurons in the perceptual engine stage, this feature should be inherited by the subsequent motion detector stage. Thus, when exposed to an oriented grating, the motion detector neurons are capable to detect this input characteristic. A motion detector neuron having tuning for a specific orientation θ will therefore encode the speed component that is orthogonal to such direction. In order to demonstrate that such orientation selectivity is actually preserved, we show in the spider chart of figure 4 the response of the motion detector neurons to variously oriented gratings: we can notice a marked tuning for the orientation feature of the grating stimulus. This ability is acquired by appropriately setting the orientation of the feedforward kernel from the sensory input to the perceptual engine stage, and by rotating accordingly the axis that aligns the two inhibitory recurrent Gaussian clusters. In particular, we have considered 12 values for the tuning orientations, evenly spaced in range $[0 - 180)$ deg, with a 30 deg step, but we actually tested the network on 24 stimulus' orientations. Such response, for any given tuning $\theta < 180$, is the average on $\mathbf{n}$ and on any positive speed $+v_f$. The curves for $\theta \geq 180$ instead were obtained by averaging the activity of neurons with opposite motion direction preference (i.e. negative speeds).

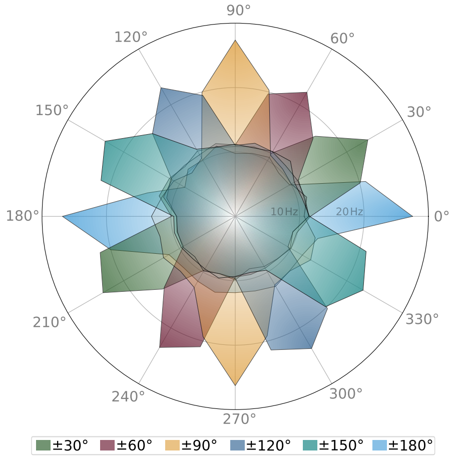

Figure 4. **Direction tuning in motion detectors.** Results from the simulations with oriented drifting gratings. Each radius represents the activity of a population at the last stage of the network, while the different colors identify the orientations of the tested stimulus. For the sake of clarity, the same colour has been used to indicate selectivity at a given stimulus' orientation and to the corresponding one with a shift of 180 deg.

Real-world application. To validate the functionality of the network, we tested it with natural stimuli, in particular a shaking drumstick. The gesture of a drum player musician moving the sticks can be characterised by different speeds and different inclinations of them. Having acquired the simulation data from the motion detector stage, the last processing stage is represented by the decoder, from which we were able to extrapolate the optic flow. The result is shown in figure 5, in which we provide a frame, selected from a video, of the event-based optic flow computation. The green arrows identify the drumsticks' direction and speed of motion superimposed on a frame obtained by accumulating events in a given time window (of 20 ms).

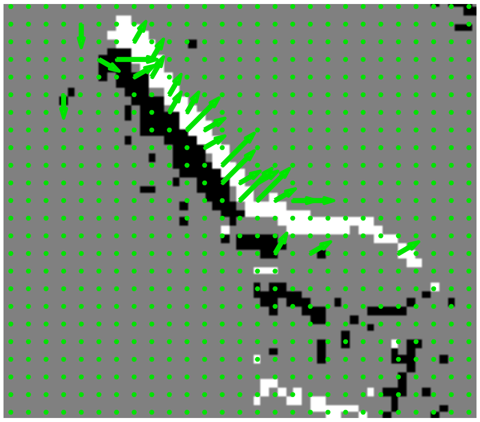

Figure 5. **Optic flow.** Optic flow from an example of real-world application. The neuromorphic sensor was placed in front of a subject moving the drumstick (in this case, upwards along a 45° direction).

6. Conclusion

In this work we have presented the architecture of spiking neural network for optic flow estimation. The network is fully bio-inspired, both at the level of cortical stages and processing, and of individual units, such as neurons and synapses. The innovative feature of our approach concerns having adapted and extended some known rate-based mechanisms to a spike-based network. Notably, we have described the application of particular synapses, having a temporal dependence, in order to obtain spatio-temporal receptive fields from combinations of static Gabor-like spatial filters. The first processing stage of our network (which we defined as the *perceptual engine*) can be seen as a basic building block. A simple but proper combination of these elements gives rise to neuronal detectors of the desired complex visual features. This architecture has therefore allowed us to obtain velocity estimates for both synthetic and natural stimuli. Our future goal is to extend the region of the processed image by considering a larger portion of the sensor input and include a multi-scale analysis that will therefore increase the computational accuracy in more complex visual scenes.

References

[1] Edward Adelson and James Bergen. The plenoptic function and the elements of early vision. 08 1997. 1

[2] Edward H. Adelson and James R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, Feb 1985. 1, 2, 3, 7

[3] Edward H. Adelson and J. Anthony Movshon. Phenomenal coherence of moving visual patterns. *Nature*, 300(5892):523–525, Dec 1982. 5

[4] Francisco Barranco, Cornelia Fermüller, and Yiannis Aloimonos. Bio-inspired motion estimation with event-driven sensors. pages 309–321, 06 2015. 2

[5] R. Benosman, Charles Clercq, Xavier Lagorce, S. Ieng, and C. Bartolozzi. Event-based visual flow. *IEEE Transactions on Neural Networks and Learning Systems*, 25:407–417, 2014. 2

[6] Ryad Benosman, Sio-Hoi Ieng, Charles Clercq, Chiara Bartolozzi, and Mandyam Srinivasan. Asynchronous frameless event-based optical flow. *Neural Networks*, 27:32–37, 2012. 2

[7] David C. Bradley and Manu S. Goyal. Velocity computation in the primate visual system. *Nature Reviews Neuroscience*, 9(9):686–695, Sep 2008. 5

[8] Kenneth H. Britten. Mechanisms of self-motion perception. *Annual Review of Neuroscience*, 31(1):389–410, 2008. PMID: 18558861. 1

[9] Tobias Brosch, Stephan Tschechne, and Heiko Neumann. On event-based optical flow detection. *Frontiers in Neuroscience*, 9:137, 2015. 2

[10] Manuela Chessa, Silvio P. Sabatini, and Fabio Solari. A systematic analysis of a v1-mt neural model for motion estimation. 173(P3), 2016. 2, 5

[11] Tobi Delbruck. Frame-free dynamic digital vision. *Proceedings of the International Symposium on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, 03 2008. 2

[12] David J. Heeger. Model for the extraction of image flow. *J. Opt. Soc. Am. A*, 4(8):1455–1471, Aug 1987. 2

[13] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 1

[14] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. 2

[15] N. V. Kartheek Medathati, Heiko Neumann, Guillaume S. Masson, and Pierre Kornprobst. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Comp. Vision and Im. Understanding*, 150:1–30, 2016. 2

[16] Guy A. Orban. Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews*, 88(1):59–89, 2008. PMID: 18195083. 1

[17] F. Paredes-Vallés, K. Y. W. Scheper, and G. C. H. E. de Croon. Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2051–2064, 2020. 2

[18] Alexandre Pouget, Kechen Zhang, Sophie Deneve, and Peter E. Latham. Statistically efficient estimation using population coding. *Neural Comp.*, 10(2):373–401, Feb 1998. 5

[19] L. Raffo, S. P. Sabatini, G. M. Bo, and G. M. Bisio. Analog vlsi circuits as physical structures for perception in early visual tasks. *IEEE Transactions on Neural Networks*, 9(6):1483–1494, 1998. 4

[20] Bodo Rueckauer and Tobi Delbruck. Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor. *Frontiers in Neuroscience*, 10, 04 2016. 2

[21] Nicole C. Rust, Valerio Mante, Eero P. Simoncelli, and J. Anthony Movshon. How mt cells analyze the motion of visual patterns. *Nature Neuroscience*, 9(11):1421–1431, Nov 2006. 2

[22] Silvio P. Sabatini. Recurrent inhibition and clustered connectivity as a basis for gabor-like receptive fields in the visual cortex. *Biological Cybernetics*, 74(3):189–202, Mar 1996. 3, 6

[23] Eero P. Simoncelli and David J. Heeger. A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743–761, 1998. 2

[24] Marcel Stimberg, Dan Goodman, Victor Benichoux, and Romain Brette. Equation-oriented specification of neural models for simulations. *Frontiers in neuroinf.*, 8:6, 02 2014. 5

[25] Stephan Tschechne, Roman Sailer, and Heiko Neumann. Bio-inspired optic flow from event-based neuromorphic sensor input. In Neamat El Gayar, Friedhelm Schwenker, and Cheng Suen, editors, *Artificial Neural Networks in Pattern Recognition*, pages 171–182, Cham, 2014. Springer International Publishing. 2