# Spike timing-based unsupervised learning of orientation, disparity, and motion representations in a spiking neural network

Thomas Barbier, Céline Teulière, Jochen Triesch

# Spike timing-based unsupervised learning of orientation, disparity, and motion representations in a spiking neural network*

Thomas Barbier
Université Clermont Auvergne, CNRS, SIGMA Clermont, Institut Pascal,
F-63000 Clermont-Ferrand, France
thomas.barbier@uca.fr

Céline Teulière
Université Clermont Auvergne, CNRS, SIGMA Clermont, Institut Pascal,
F-63000 Clermont-Ferrand, France
celine.teuliere@uca.fr

Jochen Triesch
Frankfurt Institute for Advanced Studies,
Frankfurt, Germany triesch@fias.uni-frankfurt.de

## Abstract

*Neuromorphic vision sensors present unique advantages over their frame based counterparts. However, unsupervised learning of efficient visual representations from their asynchronous output is still a challenge, requiring a rethinking of traditional image and video processing methods. Here we present a network of leaky integrate and fire neurons that learns representations similar to those of simple and complex cells in the primary visual cortex of mammals from the input of two event-based vision sensors. Through the combination of spike timing-dependent plasticity and homeostatic mechanisms, the network learns visual feature detectors for orientation, disparity, and motion in a fully unsupervised fashion. We validate our approach on a mobile robotic platform.*

## 1. Introduction

The development of deep neural networks (DNNs) for image and video processing has progressed rapidly in recent years, with current systems displaying impressive performance and versatility. Nevertheless, these systems are limited by the requirement of large amounts of labeled training data and poor energy efficiency. In contrast, the mammalian brain exhibits unequaled unsupervised learning capabilities at ultra-low energy requirements. A growing body of research on neuromorphic vision systems [17] is therefore trying to mimic these features of biological vision.

Event-based vision sensors [8] are a great starting point for exploring neuromorphic vision systems with brain-like learning abilities. These sensors are directly inspired by the information processing in the retina [19]. In particular, their pixels react independently to changes in light intensity, generating an asynchronous flow of "events", similar to the flow of action potentials or "spikes" produced by retinal ganglion cells, which communicate visual information from the retina to the brain. This asynchronous processing of event-based vision sensors gives them several advantages over traditional cameras such as very high temporal resolution at very low data rates. In addition, they exhibit high dynamic range and low power consumption. Furthermore, the sparse nature of the sensor's outputs avoids transmitting redundant information and maximizes the system's efficiency while minimizing computational load.

To fully harvest the benefits of these sensors, it is important that subsequent processing steps take advantage of their low-latency asynchronous output. Spiking Neural Networks (SNNs) are ideally suited for this [6, 26, 27], since they follow the same architectural principles as used by the brain, using an arrangement of comparatively simple computational units ("neurons") connected via weighted con-

*This work was sponsored by a public grant overseen by the French National Agency as part of the "Investissements d'Avenir" through the IMobS3 Laboratory of Excellence (ANR-10-LABX-0016) and the IDEX-ISITE initiative CAP 20-25 (ANR-16-IDEX-0001). Financial support was also received from Clermont Auvergne Metropole through a French Tech-Clermont Auvergne professorship. JT acknowledges support from the Johanna Quandt foundation.

nections ("synapses") and communicating asynchronously via spikes. At present, it is not yet fully understood how the connectivity of visual cortex is set up through developmental and learning processes [29], but it is clear that learning in the brain does not require explicit supervision. Therefore, we here attempt to mimic the unsupervised learning abilities of visual cortex in a neuromorphic vision system. To this end, we present a novel spiking neural network, inspired by so-called simple and complex cells of the mammalian visual cortex. To the best of our knowledge, it is the first SNN that learns complex cells via a timing-based rule from event-based visual input. We demonstrate that it learns feature detectors for local orientation, disparity, and motion from the asynchronous inputs of a pair of event-based vision sensors mounted on a mobile robotic platform without the need for any supervision. This brings us one step closer to the creation of neuromorphic vision systems that can learn in a more brain-like unsupervised fashion. Overall, our contributions are as follows: 1.) We present a novel spiking neural network that learns simple and complex cell-like receptive fields from binocular event-based input using biologically plausible spike timing-based unsupervised learning mechanisms. 2.) We demonstrate that the learned representations match biological findings and come to reflect the statistical properties of the input signals in terms of their distribution of orientation, motion, and disparity tuning. 3.) We validate our approach on a mobile robotic platform.

## 2. Related Work

The field of neuromorphic vision systems using event-based cameras has been expanding rapidly in recent years, as has the use of SNNs to solve complex vision tasks. Supervised learning techniques are used frequently in works combining event-based vision sensors with SNNs. One popular trend, especially in the case of recognition tasks, is to train a standard DNN using backpropagation and then convert the network to an SNN. This can lead to good results, though inferior to the DNN itself, but is not very versatile, scalable, or biologically plausible.

On the other hand, direct supervised training of SNNs is becoming more popular as modified backpropagation rules intended for SNNs have been developed [28, 31, 4]. Though much more convincing, this approach still requires extensive amounts of labeled training data.

Sidestepping the problem of learning, some research has pre-wired early visual feature detectors. Tschechne et al. [34] estimate the optical flow of a scene with the help of a set of spatio-temporal filters created by combining Gabor functions. Variants of the Barlow and Levick model, a direction sensitive neuronal triplet, have been employed to create a network capable of estimating optic flow [9, 5]. Orchard et al. [22] have used a network with synaptic delays to construct feature detectors sensitive to visual scene motions. Similarly, in the field of stereo vision [33], there have been attempts at creating networks to estimate binocular disparities between two event-based cameras [7, 23, **?**]. These methods present interesting biological properties and cleverly designed network structures, but they lack a decisive trait: the ability to learn. Learning allows SNNs to become much more versatile and adapt to changing environments. A particularly promising unsupervised learning mechanism that has been used successfully in SNNs is spike timing-dependent plasticity (STDP), of which different formulations exist. Often it is combined with homeostatic mechanisms to keep activity levels in desired regimes.

Masquelier et al. [20] demonstrated unsupervised learning of hierarchical visual representations with STDP. Akolkar et al. [1] showed the ability of STDP to learn an efficient visual representations that even surpasses Gabor filters. Kheradpisheh et al. [11] used a deep spiking neural network with STDP-based learning to solve an object recognition task with results close to supervised approaches. Srinivan et al. [32] introduced a novel "STDP-based convolution-over-time learning methodology" and apply it to hand-written digit recognition. Hopkins et al. [10] created a biologically inspired STDP-based network on the SpiNNaker neuromorphic chip and successfully tackled a recognition task. Paulun et al. [25] used a bio-inspired down-sampling of the visual field fed to a complex brain-like Neucube architecture. The learning was done with a combination of STDP rules and a last stage was comprised of a supervised SNN classifier. Lagorce et al. [13] presented an interesting unsupervised learning of visual representations, but their implementation is based on a conversion to a rate code and does not use STDP. Liu et al. [16] introduced a new "Multiscale Spatio-Temporal Feature" representation and applied it to recognition tasks such as gesture or digit recognition via STDP. Recently, Paredes-Valles et al. [24] have presented a multi-stage SNN that is capable of optical flow estimation using STDP-based learning combined with homeostatic mechanisms.

Our network model shares architectural features with several of the works above, but appears to be the first that learns complex cells with a timing-based plasticity rule, demonstrating the approach on inputs of an event-based vision sensor.

## 3. Methods

### 3.1. Leaky Integrate and Fire Neuron

We use the Leaky Integrate and Fire (LIF) neuron model as the basic building block of our model. It is one of the most popular neuron models and offers great simplicity compared to more detailed bio-physical models [18], while still providing an accurate depiction of a biological principles. An LIF neuron is described by a membrane potential

$V(t)$ that decays exponentially to a resting value over time when no inputs are registered. Here, we use a resting value $V_{\text{rest}}$ of 0 for simplicity. The exponential decay speed is controlled by the membrane time constant $\tau_m$. LIF neurons, just like real neurons, are linked via weighted connections called synapses. A synapse transmits information from one neuron to another. The strength of a synapse $w_i$ is plastic, i.e., it can change with time. Modifying the synaptic weights is the main way to control the behavior of a spiking neural network. When an LIF neuron receives an input from a synapse (called a "pre-synaptic" input) at a time $t$, its membrane potential changes by the amount $w_i$. Let $\Delta t$ be the time between the current input and the previous input to the neuron. Then the new membrane potential can be calculated directly upon arrival of the input as:

$$\tilde{V}(t+\Delta t) = V(t)\, e^{\frac{-\Delta t}{\tau_m}} + w_i(t) \tag{1}$$

If the membrane potential exceeds a threshold $V_\theta$, the neuron is said to "spike". It generates an action potential, which is then propagated to other neurons via synapses. Its membrane potential is then reset to $V_{\text{rest}} = 0$:

$$V(t+\Delta t) = \begin{cases} \tilde{V}(t+\Delta t) : \tilde{V}(t+\Delta t) < V_\theta \\ V_{\text{rest}} : \tilde{V}(t+\Delta t) \geq V_\theta \ . \end{cases} \tag{2}$$

## 3.2. Homeostatic mechanisms

The LIF neuron presented before is one of the simplest models. But biological neurons exhibit much more complex mechanisms to adapt their behavior to varying situations. We replicate some of the most beneficial of those mechanisms in order to improve the stability, adaptability and robustness of our spiking neural network.

### 3.2.1 Refractory period

When a neuron spikes, it enters in a period of low excitability called a refractory period. This strongly limits the spike frequency of neurons in the case of frequent and/or large inputs. We model a refractory mechanism through a trace $\eta_{\text{RP}}$ generated after each spike and then decaying exponentially back to zero at a rate defined by $\tau_{\text{RP}}$. The membrane potential update becomes:

$$\tilde{V}(t+\Delta t) = V(t)\, e^{\frac{-\Delta t}{\tau_m}} - \eta_{\text{RP}}\, e^{-\frac{t+\Delta t-t_s}{\tau_{\text{RP}}}}, \tag{3}$$

where $t_s$ is the time of the neuron's last spike.

### 3.2.2 Threshold and spike rate adaptation

Event-based cameras present an inherent variability in the output frequency depending on factors such as lighting conditions, the amount of textures or the relative speed of objects. This means that neurons will be subjected to variable data rate during their operation, which in turn can lead to high variability in their spike rates. Although this is somewhat unavoidable, it is preferable to keep the spike rate of neurons in a reasonable range across different conditions. Biology handles that problem by means of a wide range of homeostasis mechanisms, be it at the heart of the neuron, the soma, or at the sites of its inputs, the synapses.

We implemented two such mechanisms. The first one is oriented towards long term regulation. It constantly balances the value of the threshold $V_\theta$ depending on the spiking activity $S(t)$ of a neuron in order to reach a target spike rate $S^*$. It can be written as:

$$\Delta V_\theta = \eta_{\text{TA}}\, (S(t) - S^*)\ , \tag{4}$$

with $\eta_{\text{TA}}$ controlling the speed at which the threshold adapts. $S(t)$ is computed by counting the number of spikes which occurred in the previous 10 seconds. The threshold update happens every second, and is therefore a somewhat slow process intended to handle global illumination and speed conditions. We define a minimum threshold $V_{\theta\,\min}$ to avoid capturing camera noise in areas with very little inputs.

For local variations, we use a faster process called spike rate adaptation. Just like the threshold adaptation, it relies on the neuron's activity. When a neuron spikes, a trace $V_{\text{SRA}}(t)$ is increased by a value $\eta_{\text{SRA}}$. This trace is subtracted for each pre-synaptic input and decays exponentially back to 0 according to the parameter $\tau_{\text{SRA}}$. The membrane potential internal update becomes:

$$\tilde{V}(t+\Delta t) = V(t)\, e^{\frac{-\Delta t}{\tau_m}} - V_{\text{SRA}}(t)\, e^{\frac{-\Delta t}{\tau_{\text{SRA}}}} - \eta_{\text{RP}}\, e^{\frac{t_s-t-\Delta t}{\tau_{\text{RP}}}} \tag{5}$$

This regulation mechanism acts immediately after a spike has happened. By choosing a relatively small time constant $\tau_{\text{SRA}}$ (in the order of a few hundreds of milliseconds), the effects of the spike rate adaptation mechanism are only visible on short periods of time, which complements the slower threshold update.

### 3.2.3 Lateral inhibition

Next to excitatory synapses, our network also uses lateral inhibitory synaptic connections. However, we simplified the mechanism of inhibition compared to biology. Rather than assigning specific neurons to the role of inhibitory cells, we simply considered that some neurons can directly inhibit each other reciprocally. Specifically, when a neuron spikes, it will instantly inhibit its neighbors by subtracting a fixed value $\eta_{\text{INH}}$ from their membrane potential. This mechanism introduces strong competition between neighboring neurons.

### 3.3. Spike Timing-Dependent Plasticity

Spike timing-dependent plasticity (STDP) is a fundamental learning mechanism of the nervous system. If a neuron's spike causes another neuron to also spike, then the synaptic connection from the first neuron to the second one will be strengthened. Conversely, if the post-synaptic neuron spikes before the pre-synaptic one, this generally leads to weakening of the synaptic connection. This simple mechanism has given rise to one of the most popular techniques for unsupervised learning in SNNs via so-called STDP rules [3]. There are many different formulations of STDP, but the main idea stays the same. Synaptic weights evolve depending on the relative timing between pre and post-synaptic spikes. Long-Term Potentiation (LTP) happens when a pre-synaptic spike arrives before a post-synaptic one. The opposite leads to Long Term Depression (LTD).

In our network, the update of weights happens directly after the spiking of a neuron. We keep track of the exact timing of every pre-synaptic spike $t_i$ and the 2 last post-synaptic spikes $t_s$ and $t_{s-1}$. The update of the synaptic weights can be written as:

$$\begin{aligned} \Delta w_i^{\mathrm{LTP}} &= \eta_{\mathrm{LTP}}\, e^{\frac{t_i - t_s}{\tau_{\mathrm{LTP}}}} \\ \Delta w_i^{\mathrm{LTD}} &= -\eta_{\mathrm{LTD}}\, e^{\frac{t_{s-1} - t_i}{\tau_{\mathrm{LTD}}}} \ , \end{aligned} \tag{6}$$

where $t_s \geq t_i \geq t_{s-1}$. $\eta_{\mathrm{LTP}}$ and $\eta_{\mathrm{LTD}}$ are parameters controlling the height of the potentiation and depression windows, whereas $\tau_{\mathrm{LTP}}$ and $\tau_{\mathrm{LTD}}$ control the width of these windows.

#### 3.3.1 Weight normalization

The STDP rule above does not include a weight limitation mechanism. To improve stability and avoid unbounded growth of synaptic weights, we force the weights to be positive and use a separate weight normalization where synapses from inputs of each event polarity (on and off) is normalised separately to sum to a target value.

#### 3.3.2 Parallel synapses with different delays

As a final mechanism, we allow a pixel from the sensor array to connect to a neuron in the network via multiple "parallel" synapses with different transmission delays. This enables such neurons to learn a representation of optic flow via STDP.

### 3.4. Spiking neural network architecture

#### 3.4.1 Event-based pixel array

The event-based camera pixel array serves as the input layer of the SNN. Event-based cameras record the polarity of events, indicating the sign of the of change in light intensity. We therefore separate the events in 2 maps depending on their polarity (on and off) as depicted in Fig. 1.

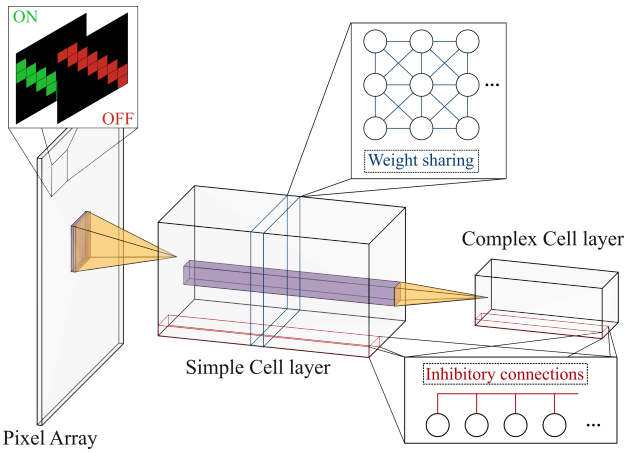



Figure 1: Proposed SNN architecture

#### 3.4.2 Simple cell layer

Found in the human early visual pathway, simple cells are neurons local orientation information. They typically form receptive fields in the shape of Gabor functions tuned to a specific orientation and possibly motion direction. The first layer of our network aims to mimic the behavior of these simple cells. It is fed directly from the pixel array. Table 1 presents the parameters used during training. We chose parameters that are mostly plausible with regard to biology.

To improve the diversity of learned receptive fields, we used a weight sharing mechanism between simple cells. Neurons looking at different locations of the visual field jointly learn the same set of synaptic weights. We define these neurons as belonging to the same "neuronal map". This is represented by the blue cross section in the simple cell layer in Fig. 1. To obtain simple cells tuned to different orientations, we increase the number of neuronal maps. We also connect neurons looking at the same location of the visual field, but from different neuronal maps, with inhibitory connections. This is represented by the red inhibitory connections in Fig. 1.

#### 3.4.3 Complex cell layer

Neurons in the complex cell layer receive inputs from simple cells. They pool inputs from a larger portion of the visual field. Similar to complex cells in visual cortex, they should learn to represent to oriented edges independently of the precise location of the edge, which requires a strongly non-linear behavior [15] [12]. Often, complex cell-like behavior is achieved via a max-pooling operation, but here we

*Simple cells:*

| $\eta_{\text{LTP}}$ (mV) | $\eta_{\text{LTD}}$ (mV) | $\eta_{\text{TA}}$ (mV) | $\eta_{\text{SRA}}$ (mV) | $\eta_{\text{RP}}$ (mV) | $\eta_{\text{INH}}$ (mV) | $\tau_{\text{LTP}}$ (ms) | $\tau_{\text{LTD}}$ (ms) | $\tau_{\text{SRA}}$ (ms) | $\tau_{\text{RP}}$ (ms) | $\tau_m$ (ms) |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.00077 | 0.00021 | 1 | 0.6 | 1 | 25 | 7 | 14 | 100 | 20 | 18 |

| $V_\theta$ (mV) | $V_{reset}$ (mv) | $V_{\theta\,\min}$ (mV) | $S^*$ (spikes/s) | $\lambda$ | $x_{\text{RF}}$ px | $y_{\text{RF}}$ px | $z_{\text{RF}}$ pol |
|---|---|---|---|---|---|---|---|
| 20 | -20 | 5 | 0.75 | 4 | 10 | 10 | 2 |

*Complex cells:*

| $\eta_{\text{LTP}}$ (mV) | $\eta_{\text{LTD}}$ (mV) | $\eta_{\text{TA}}$ (mV) | $\eta_{\text{SRA}}$ (mV) | $\eta_{\text{RP}}$ (mV) | $\eta_{\text{INH}}$ (mV) | $\tau_{\text{LTP}}$ (ms) | $\tau_{\text{LTD}}$ (ms) | $\tau_{\text{SRA}}$ (ms) | $\tau_{\text{RP}}$ (ms) | $\tau_m$ (ms) |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 | 0.2 | 1 | 0.6 | 1 | 25 | 20 | 20 | 100 | 20 | 20 |

| $V_\theta$ (mV) | $V_{reset}$ (mv) | $V_{\theta\,\min}$ (mV) | $S^*$ (spikes/s) | $\lambda$ | $x_{\text{RF}}$ cell | $y_{\text{RF}}$ cell | $z_{\text{RF}}$ cell |
|---|---|---|---|---|---|---|---|
| 3 | -20 | 2 | 0.75 | 10 | 4 | 4 | 100 |

Table 1: Parameters used for the simple and complex cells.

are interested in learning the non-linear behavior of complex cells with STDP. For this, we adjust complex cell parameters to be more sensitive to inputs by giving them a lower spiking threshold (cf. Tab. 1).

We also chose a different and simpler STDP window for the complex cells. Contrary to simple cells, the weight variation is always positive, following a step function. Unbounded growth is avoided using weight normalization. The STDP rule for complex cells is written as:

$$\Delta w_i^{\text{LTP,c}} = \begin{cases} \eta_{\text{LTP}} : |t_i - t_s| \leq \tau_{\text{LTP}} \\ 0 : |t_i - t_s| > \tau_{\text{LTP}} \ . \end{cases}$$
$$\Delta w_i^{\text{LTD,c}} = \begin{cases} \eta_{\text{LTD}} : |t_{s-1} - t_i| \leq \tau_{\text{LTD}} \\ 0 : |t_{s-1} - t_i| > \tau_{\text{LTD}} \ . \end{cases} \quad (7)$$

We also tried other STDP windows, such as a linear or exponential ones. We compare these results in the supplementary materials, cf. suppl. Fig. 2.

## 4. Results

### 4.1. Learning simple cell receptive fields

We first evaluate the ability for our network's simple cell layer to learn 3 essential environment properties: orientation, motion, and disparity. We created a synthetic event video where we precisely controlled those 3 variables. The simplest type of stimuli consist of moving bars (from left to right) in front of a static camera. We chose a bright bar moving on a dark background, as depicted in Fig. 2a. This results in a leading positive event polarity edge, followed by a negative polarity edge. We created the artificial video by creating stereoscopic frames and then converting them to event-based videos using the ESIM simulator [30]. Because of the synthetic nature of the stimulus, we can choose precise bar orientations, motions, and disparities.

The learned simple receptive fields are shown in Figs. 2b and 2c. They are tuned to vertical orientations, matching the bars orientations in the synthetic videos. As the bars are

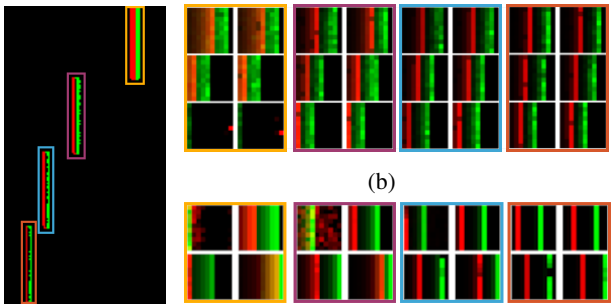
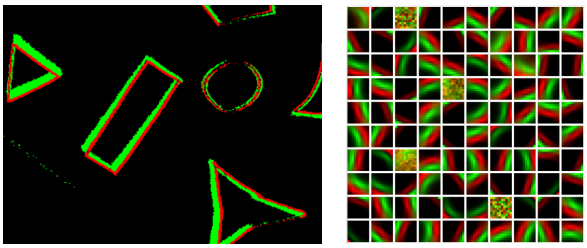



Figure 2: (a) Synthetic event video made of vertical bars moving from left to right. Their speed varies from top (fastest) to bottom (slowest). (b) Motion sensitive cells (top) with 3 increasing synaptic delays (represented as 3 squared receptive field stack on top of each other). (c) Disparity sensitive cells (bottom) connected to a left and right "synthetic" camera (represented as 2 squared receptive field stacked on top of each other). 2 neurons are presented per moving bar, from the fastest bar (left) to the slowest (right).

moving from left to right in the visual field, events of specific polarities will always appear in the same order (positive then negative). The learned receptive fields reflect this specific order and resemble Gabor functions describing biological simple cell receptive fields [2].

When adding multi-synaptic connections with multiple delays between the pixels and the simple cells, they learn a representation of the speeds of the bars. Specifically, simple cells are connected to the pixel array using 3 synaptic connections with different delays per pixel. This is represented in Fig. 2b by showing the corresponding 3 receptive subfields (one per delay) on top of one another. For each bar we show two examples of learned receptive fields (correspondence is indicated by colored frames). The four bars are moving at speeds of (top to bottom) 420, 210, 140 and 105 pixels/s relative to the camera. To accurately capture this motion, we chose synaptic delays of 0, 10 and 20 ms for receptive fields of 10 by 10 pixels. This amounts to a velocity tuning of up to $\frac{10}{2.10^{-3}} = 500\,\text{pixels/s}$. Above that speed, the receptive fields would be too small (or the delays too big) to capture the bar motion. In Fig. 2b we see that faster speeds become reflected in bigger shifts between the subfields corresponding to the different delays. For the fastest moving bar (close to the 500 pixels/s limit), the receptive fields start to "break" for the 20 ms delay (bottom row in (b). This is because the first subfield with delay zero does not start completely to the right, pushing the shift for the 20 ms synaptic delay beyond the size of the receptive field. Due to the polarity independent weight normalization, the weight to a random synaptic pixel becomes very strong (single red pixel).



(a) (b)

Figure 3: (a) Screenshot of an event video of various shapes moving in front of a DAVIS 346 camera. (b) Resulting receptive fields of simple cells learned with the moving shapes video sequence.

We extend this experiment by adding a second set of moving bars in order to form a stereoscopic setup. Each bar is slightly shifted in the second visual field to mimic different binocular disparities. Specifically, the bars have disparities of (top to bottom) 2, 4, 6, and 8 pixels. Here, simple cells are connected to the 2 stereoscopic visual fields via single synaptic connections. Fig. 2c depicts the resulting left and right receptive fields placed on top of each other. As expected, the shift between the left and right receptive subfields matches the disparity of the bar.

### 4.2. Learning complex cell receptive fields

We extend the network by adding a layer of complex cells and changing the input stimulus to a real event-based recording made with a DAVIS 346 camera. We moved a sheet of paper with various shapes drawn onto it in front of the camera. Real inputs from event-based camera contain a substantial amount of noise that can affect the training. We used an event noise filter available in the DV-software (associated with the DAVIS 346 camera) to reduce this effect. The video presents many edges in many orientations, as depicted in Fig. 3a.

To increase variability, we perform data augmentation, by presenting the same video with artificial rotations and mirroring effects during training. This ensures that the neurons are presented with edges of all possible orientations. We use a first layer of simple cells without the use of multi-synaptic connections and add a second layer of complex cells and trained the network for approximately 30 minutes.

Figure 3b presents the resulting receptive fields of the simple cells. They are also Gabor-shaped but show a wide range of orientation tuning. Four of the 100 simple cells were not able to learn anything. This is due to the strong inhibition in the network that prevented them from becoming active on a regular basis. Indeed, when we increased the number of simple cells to 144 we found that the number of cells with Gabor-shaped receptive fields stayed roughly constant, because an increasing number of cells became permanently suppressed.

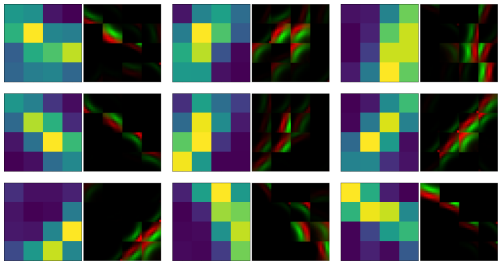
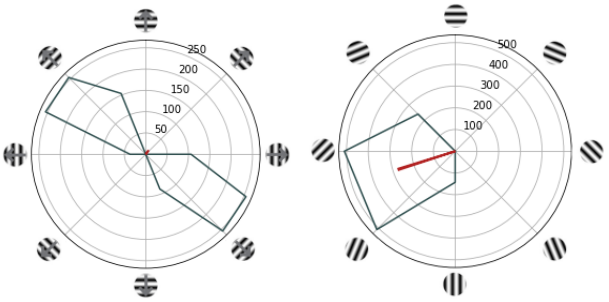


Figure 4: Visualization of the RFs of six example complex cells. Left images represent the total strengths of connections between a complex cells and simple cells from sixteen different input regions (brighter is stronger). Right images show the RFs of the simple cells that have the strongest connection to the complex cell for each of the 16 different regions. Simple cell RFs are scaled by their connection strength to the complex cell.

Complex cells' receptive field are harder to visualize since they pool activity from many simple cells connected to a larger area of the pixel array in a nonlinear fashion. Figure 4 represents the weighted connections to a complex cell from simple cells connected to it. The left images were created by summing all the weights from the simple cell maps. Yellow squares represent the strongest connections. The complex cell receptive fields have become localized, i.e., they respond to inputs from local regions of the pixel array. The images on the right are constructed by selecting the simple cell receptive field with the strongest connection to the complex cell for each of the local region of the pixel array that projects to it. Typically, these simple cell receptive fields have similar orientation but diverse phases, giving rise to phase invariance of complex cells. Furthermore, while simple cells are selective for a particular motion direction, i.e., they are direction selective, our complex cells are mostly only orientation selective.

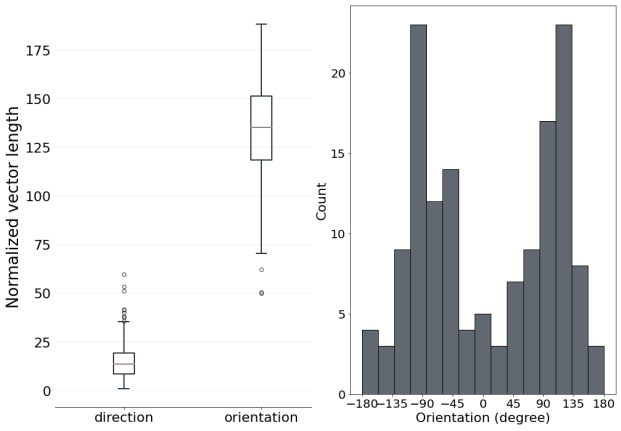
Neuroscientists usually analyse the behaviours of these types of cells by observing their responses to oriented stimuli. A standard test is to show different oriented gratings and measure a cell's response. We followed the same procedure by creating artificial event sequences as produced by gratings moving in 16 different directions (from 0 to 360 degrees by steps of 22.5 degrees). We measured the responses of our artificial neurons by counting the number spikes produced during the presentation of the stimuli and averaging the numbers over 5 trials. We used the normalized vector length $L$ to quantify orientation and direction tuning [21]:

$$\begin{aligned} L_{\text{dir}} &= \left| \frac{\sum_k R(\theta_k) \exp(i\theta_k)}{\sum_k R(\theta_k)} \right| \\ L_{\text{ori}} &= \left| \frac{\sum_k R(\theta_k) \exp(2i\theta_k)}{\sum_k R(\theta_k)} \right| , \end{aligned} \tag{8}$$




(a) (b)

Figure 5: (b) Complex cell response in direction space, made from counting the cell's spikes for a rotating grating stimulus. The red line corresponds to the normalized vector length and indicates the cell's selectivity strength and direction. (c) Complex cell response in orientation space, made by pooling over opposite directions.




(a) (b)

Figure 6: (a) Normalized vector length distribution of the network complex cells in direction and orientation space. (b) Histogram of normalized vector orientations in orientation space (0° corresponds to a horizontal orientation).

where $R(\theta_k)$ is the average number of spikes for stimulus direction $\theta_k$. For orientations, we simply summed together the number of spikes for the 2 opposite motion directions. The normalized vector length gives the sensitivity of a cell to a particular direction. A cell that reacts very strongly to only one direction is said to be highly direction selective. Whereas a cell reacting strongly to many different directions is not selective. A particular case emerges with cells being very selective in orientation space but not in direction space. This is due to cells responding strongly to opposing motion directions.

We visualize a cell's selectivity by plotting a circular histogram of its responses to different oriented stimuli. Figure 5 shows the response of an example complex cell in orientation and direction space. The red lines correspond to the normalized vector length. We observe that this cell is orientation selective 5b but not direction selective 5a. This is because it exhibits 2 roughly symmetric lobes in direction space, which cause a low selectivity value. The cell will strongly respond to stimuli oriented at around 135° and 315°, but not to other orientations. Looking at the full set of the 144 learned complex cells in direction space, we observe that most complex cells exhibit similar responses to the ones shown here, except for a few that are also direction selective (uni-lobe). We present a larger set of 36 complex cells in the supplementary material, suppl. Fig. 1.

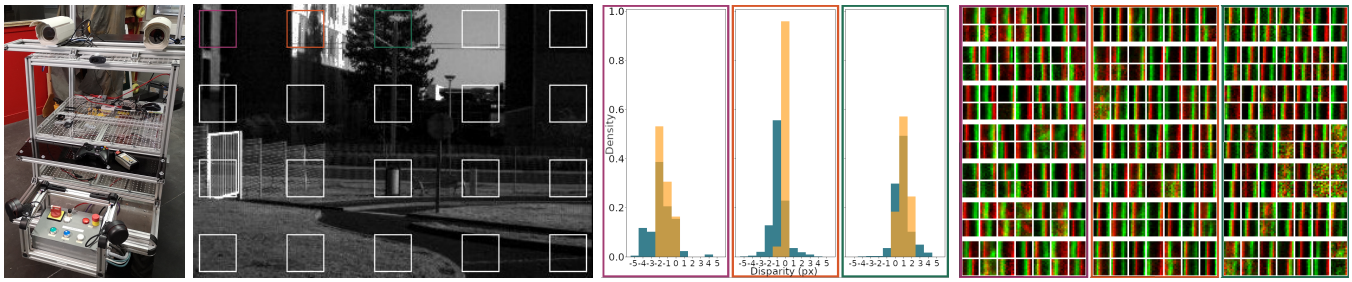
To quantify the results, we compute the normalized vector length for the full batch of trained complex cells (144 in total). The result is visualized in Fig. 6. It shows the overall orientation and direction selectivity. On average, cells are highly orientation selective but not very direction selective. Figure 6b represents the distribution of preferred orientations of the complex cells. It exhibits a preference for oblique orientation. This is consistent with a slight overall preference for oblique orientations among simple cells seen in Fig. 3b. Overall, the set of complex cells has learned to respond to a wide range of orientations/motion directions.

### 4.3. Stereo driving scenes

In this section, we feed our network stereoscopic input from a pair of event-based cameras mounted on a mobile robotic platform operating in an urban outdoor environment (see Fig. 7a). The disparity statistics vary greatly depending on the location within the visual field. We were interested to see whether these statistical difference become reflected in the network's receptive fields. We use a network composed of multiple patches of neurons looking at specific regions of the visual field. Neurons in different locations within a single region share their weights (similar to convolutional neural networks). Specifically, we consider 20 regions, each one composed of an array of 3 by 3 simple cells, with 49 neuronal maps connected by inhibitory synapses. Figure 7b shows a typical image from the training sequence. The 20 regions are marked by squares.

The horizontal distance between the two event-based sensors induces disparities between the two images. Neurons are connected to the inputs of both the left and right sensor and receive related events from both, but at slightly shifted pixel locations depending on the distance to the stimuli. This induces differences between the learned left and right receptive subfields of such a binocular neuron. We train the network on a small repeated sequence of about 45 seconds to control the number of different disparities to which the neurons are exposed.

We find the preferred disparity of neuron's learned receptive field by determining the smallest mean squared er-




(a) (b) (c) (d)

Figure 7: (a) Mobile platform with the pair of stereo event-based cameras mounted on top. (b) Example of an image (left camera) taken from one of the outdoor sequences. The squares mark the different image regions. The colored squares in the upper left mark regions for which disparities and receptive fields are shown in (c) and (d). (c) Histograms of estimated disparities of the learned receptive fields (orange) and disparities estimated from corresponding image frames with conventional computer vision techniques (blue) for the 3 colored regions. (d) Learned left (top) and right (bottom) receptive subfields for the 49 neuron layers in the 3 colored regions.

ror (MSE) between the left and shifted versions of the right receptive subfields of the neuron. It is important to note that the maximum possible disparity is limited by the size of the receptive fields. The learned receptive fields are shown in Fig. 7d. We selected 3 regions that were mostly exposed to objects of similar distances so that input to each region was dominated by a different disparity. We observe that only vertical receptive fields were learned for this input. This is likely due to the scene structure, which is dominated by vertical orientations. More generally, it is well-known that vertical and horizontal orientations are abundant in man-made environments [14]. Figure 7c presents a histogram (orange) of the computed disparities. We observe that most receptive fields in a region learned one specific disparity, up to a variation of 1 or 2 pixels. Interestingly, there is a systematic relation across the regions. The learned disparity increases roughly linearly as we move from the left side to the center of the visual field. This is consistent with the structure of the scene, where objects in the center are closer. We compare the learned preferred disparities to disparities estimated via conventional computer vision techniques from image frames (that the sensor also produces) (blue histogram in Fig. 7c). The results are in reasonable agreement with the receptive fields learned by our spiking network.

We also trained our network on the MVSEC dataset [35], focusing on the sequence "outdoorday1data", a stereo urban driving scene. RFs are learned for nine different regions of the visual field (similar to the 20 regions considered in Fig. 7). The learned preferred disparities for the different regions reflect different distribution of object distances across these regions roughly corresponding to the ground truth data of object distances, cf. suppl. Fig. 3. Learned simple cell RFs are shown in suppl. Fig. 4. The set of RFs forms a strong basis responding to multiple orientations and disparities consistent with the scene structure, making it an efficient processing stage for solving more complex dynamic tasks.

## 5. Discussion

Taking inspiration from neurobiology, we have presented a spiking neural network that learns orientation, motion, and disparity representations in a fully unsupervised fashion via STDP from input from a stereoscopic event-based vision setup mounted on a mobile robotic platform. The learned representation shares many similarities to that observed in the brain, including simple and complex cells as found in primary visual cortex of mammals. Furthermore, as observed in biology, the learned representation adapts to the statistics of the visual input (in terms of orientation, motion, and disparity). From a biological standpoint, we have made a number of gross simplifications including the instantaneous inhibition or simple weight normalization. Therefore, our work should not be seen as an attempt to construct a faithful model of biological learning. However, being fully spike-based and relying on only local learning rules, our network is well-suited for implementation on modern neuromorphic spiking network hardware. This will facilitate scaling up our approach to much larger network sizes. This is left for future work. Furthermore, in the future we would like to extend this work to active binocular and motion vision and develop techniques for the autonomous self-calibration of spike-based active binocular vision systems.

## References

[1] Himanshu Akolkar, Stefano Panzeri, and Chiara Bartolozzi. Spike time based unsupervised learning of receptive fields for event-driven vision. pages 4258–4264, 2015. 2

[2] Thomas Barbier, Céline Teulière, and Jochen Triesch. Unsupervised Learning of Spatio-Temporal Receptive Fields from an Event-Based Vision Sensor. pages 622–633, 2020. 5

[3] Natalia Caporale and Yang Dan. Spike timing–dependent plasticity: A hebbian learning rule. *Annual Review of Neuroscience*, 31(1):25–46, 2008. 4

[4] Xiang Cheng, Yunzhe Hao, Jiaming Xu, and Bo Xu. LISNN: Improving spiking neural networks with lateral interactions for robust object recognition. 1:1519–1525, 2020. 2

[5] Giulia D'Angelo, Ella Janotte, Thorben Schoepe, James O'Keeffe, Moritz B. Milde, Elisabetta Chicca, and Chiara Bartolozzi. Event-Based Eccentric Motion Detection Exploiting Time Difference Encoding. *Frontiers in Neuroscience*, 14:451, 2020. 2

[6] Peter U. Diehl and Matthew Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, 9:99, 2015. 1

[7] Georgi Dikov, Mohsen Firouzi, Florian Röhrbein, Jörg Conradt, and Christoph Richter. Spiking Cooperative Stereo-Matching at 2 ms Latency with Neuromorphic Hardware. pages 119–137, 2017. 2

[8] G. Gallego, T. Delbruck, G. M. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1

[9] Germain Haessig, Andrew Cassidy, Rodrigo Alvarez, Ryad Benosman, and Garrick Orchard. Spiking Optical Flow for Event-Based Sensors Using IBM's TrueNorth Neurosynaptic System. *IEEE Transactions on Biomedical Circuits and Systems*, 12(4):860–870, 2018. 2

[10] Michael Hopkins, Garibaldi Pineda-Garcia, Petrut A. Bogdan, and Steve B. Furber. Spiking neural networks for computer vision. *Interface Focus*, 2018. 2

[11] Saeed Reza Kheradpisheh, Mohammad Ganjtabesh, Simon J. Thorpe, and Timothée Masquelier. STDP-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, 99:56–67, 2018. 2

[12] Konrad P. Körding, Christoph Kayser, Wolfgang Einhäuser, and Peter König. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 91(1):206–212, 2004. 4

[13] Xavier Lagorce, Sio Hoi Ieng, Xavier Clady, Michael Pfeiffer, and Ryad B. Benosman. Spatiotemporal features for asynchronous event-based data. *Frontiers in Neuroscience*, 9, 2015. 2

[14] Baowang Li, Matthew R Peterson, and Ralph D Freeman. Oblique effect: a neural basis in the visual cortex. *Journal of neurophysiology*, 90(1):204–217, 2003. 8

[15] Yanbo Lian, Ali Almasi, David B. Grayden, Tatiana Kameneva, Anthony N. Burkitt, and Hamish Meffin. Learning receptive field properties of complex cells in v1. *bioRxiv*, 2020. 4

[16] Qianhui Liu, Gang Pan, Haibo Ruan, Dong Xing, Qi Xu, and Huajin Tang. Unsupervised AER Object Recognition Based on Multiscale Spatio-Temporal Features and Spiking Neurons. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5300–5311, 2020. 2

[17] Shih Chii Liu, Tobi Delbruck, Giacomo Indiveri, Adrian Whatley, and Rodney Douglas. *Event-based neuromorphic systems*. 2014. 1

[18] Lyle N. Long and Guoliang Fang. A review of biologically plausible neuron models for spiking neural networks. *AIAA Infotech at Aerospace 2010*, 2010. 2

[19] Richard H. Masland. The Neuronal Organization of the Retina. *Neuron*, 76(2):266–280, 2012. 1

[20] Timothée Masquelier and Simon J Thorpe. Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity. *PLoS Computational Biology*, 3(2):31, 2007. 2

[21] Mark Mazurek, Marisa Kager, and Stephen D. Van Hooser. Robust quantification of orientation selectivity and direction selectivity. *Frontiers in Neural Circuits*, 8:92, 2014. 6

[22] Garrick Orchard, Ryad Benosman, Ralph Etienne-Cummings, and Nitish V. Thakor. A spiking neural network architecture for visual motion estimation. pages 298–301, 2013. 2

[23] Marc Osswald, Sio Hoi Ieng, Ryad Benosman, and Giacomo Indiveri. A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems. *Scientific Reports*, 7, 2017. 2

[24] Federico Paredes-Valles, Kirk Yannick Willehm Scheper, and Guido Cornelis Henricus Eugene 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*, pages 1–1, 2019. 2

[25] Lukas Paulun, Anne Wendt, and Nikola Kasabov. A retinotopic spiking neural network system for accurate recognition of moving objects using neucube and dynamic vision sensors. *Frontiers in Computational Neuroscience*, 12:42, 2018. 2

[26] J. A. Pérez-Carrasco, C. Serrano, B. Acha, T. Serrano-Gotarredona, and B. Linares-Barranco. Spike-based convolutional network for real-time processing. pages 3085–3088, 2010. 1

[27] Michael Pfeiffer and Thomas Pfeil. Deep Learning With Spiking Neurons: Opportunities and Challenges. *Frontiers in Neuroscience*, 12:774, 2018. 1

[28] Filip Ponulak, F Ponulak, and A Kasí nski. Supervised Learning in Spiking Neural Networks with ReSuMe: Sequence Learning, Classification, and Spike Shifting. 22:467–510, 2010. 2

[29] Karine Pozo and Yukiko Goda. Unraveling mechanisms of homeostatic synaptic plasticity, 2010. 2

[30] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. 2018. 5

[31] Sumit Bam Shrestha and Garrick Orchard. SLAYER: Spike Layer Error Reassignment in Time. 31, 2018. 2

[32] Gopalakrishnan Srinivasan, Priyadarshini Panda, and Kaushik Roy. STDP-based unsupervised feature learning using convolution-over-time in spiking neural networks for energy-efficient neuromorphic computing. *ACM Journal on Emerging Technologies in Computing Systems*, 14(4):1–12, 2018. 2

[33] Lea Steffen, Daniel Reichard, Jakob Weinland, Jacques Kaiser, Arne Roennau, and Rüdiger Dillmann. Neuromorphic stereo vision: A survey of bio-inspired sensors and algorithms. *Frontiers in Neurorobotics*, 13:1–20, 2019. 2

[34] Stephan Tschechne, Roman Sailer, and Heiko Neumann. Bio-inspired optic flow from event-based neuromorphic sensor input. 8774:171–182, 2014. 2

[35] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 8