Embodied Neuromorphic Vision with Event-Driven Random Backpropagation

Jacques Kaiser[1], Alexander Friedrich[1], J. Camilo Vasquez Tieck[1],
Daniel Reichard[1], Arne Roennau[1], Emre Neftci[2], Rüdiger Dillmann[1]

***Abstract*—Spike-based communication between biological neurons is sparse and unreliable. This enables the brain to process visual information from the eyes efficiently. Taking inspiration from biology, artificial spiking neural networks coupled with silicon retinas attempt to model these computations. Recent findings in machine learning allowed the derivation of a family of powerful synaptic plasticity rules approximating backpropagation for spiking networks. Are these rules capable of processing real-world visual sensory data? In this paper, we evaluate the performance of Event-Driven Random Back-Propagation (eRBP) at learning representations from event streams provided by a Dynamic Vision Sensor (DVS). First, we show that eRBP matches state-of-the-art performance on the DvsGesture dataset with the addition of a simple covert attention mechanism. By remapping visual receptive fields relatively to the center of the motion, this attention mechanism provides translation invariance at low computational cost compared to convolutions. Second, we successfully integrate eRBP in a real robotic setup, where a robotic arm grasps objects according to detected visual affordances. In this setup, visual information is actively sensed by a DVS mounted on a robotic head performing microsaccadic eye movements. We show that our method classifies affordances within 100ms after microsaccade onset, which is comparable to human performance reported in behavioral study. Our results suggest that advances in neuromorphic technology and plasticity rules enable the development of autonomous robots operating at high speed and low energy consumption.**

***Index Terms*—Neurorobotics, Spiking Neural Networks, Event-Based Vision, Learning, Synaptic Plasticity**

I. Introduction

THERE are important discrepancies between the computational paradigms of the brain and modern computer architectures. Remarkably, evolution discovered a way to learn and perform energy-efficient computations from large networks of neurons. Particularly, the communication of spikes between neurons is asynchronous, sparse and unreliable. Understanding what computations the brain performs and how they are implemented on a neural substrate is a global endeavor of our time. From an engineering perspective, this would enable the design of autonomous learning robots operating at high speed for a fraction of the energy budget of current solutions.

Recently, a family of synaptic learning rules building on groundbreaking machine learning research have been proposed in [1]–[4]. These rules implement variations of Back-Propagation (BP) adapted to spiking networks by approximating the gradient computations. As was shown in [5], [6], approximations of the gradient can be made without significantly deteriorating the accuracy of the network. These approximations allow weight updates to only depend on local information available at the synapse, enabling online learning with an efficient neuromorphic hardware implementation. Are these rules capable of efficiently processing complex visual information like the brain? Synaptic learning rules are rarely evaluated in an embodiment, and often report below state-of-the-art performance compared to classical methods [7].

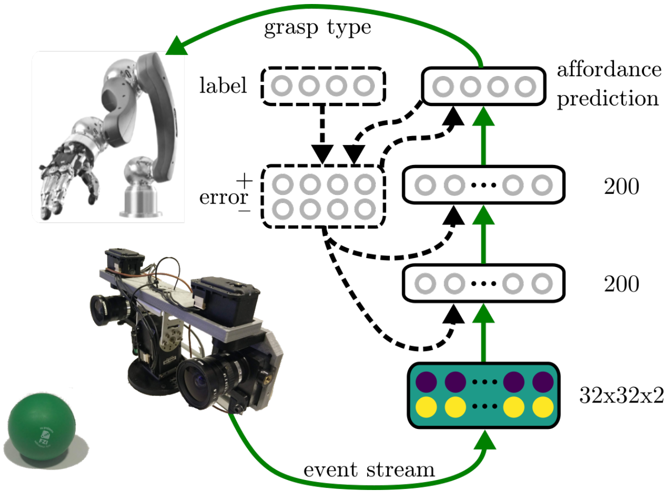


Figure 1: Our robotic setup embodying the synaptic learning rule eRBP [1]. The DVS is mounted on a robotic head performing microsaccadic eye movements. The spiking network is trained (dashed-line connections) in a supervised fashion to classify visual affordances from the event streams online. During training, output neurons and label neurons project to an error neuron population, which conveys a random feedback to hidden layers. The output neurons of the network correspond to the four types of affordances: ball-grasp, bottle-grasp, pen-grasp or do nothing. At test time, a Schunk LWA4P arm equipped with a five-finger Schunk SVH gripper performs the detected reaching and grasping motion.

In this paper, we evaluate the ability of one of these rules – eRBP [1] – to learn from real visual event-based sensor data. eRBP was the first synaptic plasticity rule formally derived from Random Backpropagation [5], its deep learning equivalent. First, we benchmark eRBP on the popular visual event-based dataset DvsGesture [8] from IBM. We introduce a covert attention mechanism which further improves the performance by providing translation invariance without convolu-

[1] FZI Research Center For Information Technology, Karlsruhe, Germany
[2] Department of Cognitive Sciences, University of California Irvine, Irvine, USA

tions. Second, we integrate the rule in a real-world closed-loop robotic grasping setup involving a robotic head, arm and a five-finger hand. The spiking network learns to classify different types of affordances based on visual information obtained with microsaccades, and communicates this information to the arm for grasping. This real-world task has the potential to enhance neuromorphic and neurorobotics research since many functional components such as reaching, grasping and depth perception can be easily segregated and implemented with brain models. This work paves the way towards the integration of brain-inspired computational paradigms into the field of robotics.

This paper is structured as follows. We give a brief introduction to the synaptic learning rule eRBP in Section II together with the network architecture. We further present our covert attention mechanism and the microsaccadic eye movements. Our approach is evaluated in Section III, first on the DvsGesture [8] benchmark, second in our real-world grasping setup. We conclude in Section IV.

II. Approach

A. Event-Driven Random Backpropagation

eRBP [1] is an interpretation for spiking neurons of feedback alignment presented in [5] for analog neurons. Feedback alignment is an approximation of backpropagation, where the prediction error is fed back to all neurons directly with fixed, random feedback weights. With a mean-square error loss, the update for a hidden weight from neuron j to neuron i can be formulated as follows:

$$\begin{aligned} \Delta w_{ij}(t) &= y_j \phi' \left(\sum_{j'} w_{ij'}(t) y_{j'}(t) \right) T_i(t), \\ T_i(t) &= \sum_k e_k(t) g_{ik}, \end{aligned} \tag{1}$$

with y_j the output of neuron j, ϕ the activation function of neuron i, e_k the prediction error for output neuron k and g_{ik} a fixed random feedback weight. This rule is interpreted for spiking neurons with y_j the spiking output (0 or 1) of neuron j and ϕ the function mapping membrane potential to spikes. The weighted sum of input spikes $\sum_{j'} w_{ij'}(t) y_{j'}(t)$ is interpreted as the membrane potential I_i of neuron i. eRBP relies on hard-threshold Leaky Integrate-And-Fire neurons, leading ϕ to be non-differentiable. Following the approach of surrogate gradients [9], ϕ' is approximated as the boxcar function, equal to 1 between b_{min} and b_{max} and 0 otherwise. Note that the temporal dynamics of the Leaky Integrate-And-Fire neuron – post-synaptic potentials and refractory period – are not taken into account in Equation (1). More recent plasticity rules now include an eligibility trace [2]–[4] accounting for some of these dynamics.

This rule can be efficiently implemented in a spiking network with weight updates triggered by pre-synaptic spikes y_j as:

$$\Delta w_{ij}(t) \propto \begin{cases} -\sum_k e_k(t) g_{ik} & \text{if } b_{\text{min}} < I_i < b_{\text{max}} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

with b_{min} and b_{max} the window of the boxcar function. The weight update for the last layer is an exception, since there are no random feedback weights. Each time a pre-synaptic neuron j with a connection to an output neuron k spikes, the weight update of this connection is calculated by:

$$\Delta w_{kj}(t) \propto \begin{cases} -e_k(t) & \text{if } b_{\text{min}} < I_k < b_{\text{max}} \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Since eRBP only uses two comparisons and one addition for each pre-synaptic spike to perform the weight update, it allows a real-time, energy-efficient and online learning implementation running on neuromorphic hardware.

B. Network Architecture

We propose a similar network architecture to the one proposed in [1]. Specifically, the network consists of spiking neurons organized in feedforward layers, performing classification with one-hot encoding. In other words, the k output neurons of the last layer correspond to the k class of the dataset. The error signals $e_k(t)$ are encoded in spikes and provided to all hidden neurons with backward connections from error neurons. These error spikes are integrated in a dedicated leaking dendritic compartment locally for every learning neuron. The weight of the backward connections are drawn from a random distribution and fixed – they correspond to the factors g_{ik} in Equations (1) and (2).

For each class, there are two error neurons conveying positive and negative error respectively. This is required, since spikes are not signed events. A pair of positive and negative error neurons are connected to the corresponding output neuron and label neuron. There is one label neuron for each class, which is spiking repeatedly during training when a sample of the respective class is presented. Formally, the error $e_k(t)$ is approximated as:

$$\begin{aligned} e_k(t) &\cong \nu_k^+(t) - \nu_k^-(t), \\ \nu_k^+(t) &\propto \nu_k^P(t) - \nu_k^L(t), \\ \nu_k^-(t) &\propto -\nu_k^P(t) + \nu_k^L(t), \end{aligned} \tag{4}$$

where $\nu_k^P(t), \nu_k^L(t)$ are the firing rates of the output neurons and label neurons respectively, and $\nu_k^+(t), \nu_k^-(t)$ are the firing rates of the positive and negative error neurons. Within this framework, all computations are performed with spiking neurons and communicated as spikes, including the computation of errors. As in the brain, all synapses are stochastic, with a probability of dropping spikes. The network is depicted in Figure 1.

In this work, the network learns from event streams provided by a DVS. Since spikes are not signed events, we associate two neurons for each pixel to convey ON- and OFF-events separately. This distinction is important since event polarities carry the instantaneous information of direction of motion (see Figure 4). Since events are generated only upon light change, two different setup are analyzed: a dataset where changes originate from motion in the scene, and a dataset where changes originate from fixational eye movements. The evaluation on these two types of dataset is important since they can lead to different performance [3].

C. Covert Attention Window

It was shown in biology that visual receptive fields of frontal eye field neurons are constantly remapped [10], [11]. Inspired from this insight, we introduce a simple covert attention mechanism which consists of moving an attention window across the input stream. Covert attention, as opposed to overt attention, signifies an attention shift which was not marked by eye movements. Particularly suited to event streams, the center of the attention window is computed online as the median address event of the last $n_{\text{attention}}$ events, see Figure 4. By remapping receptive fields relatively to the center of the motion, this technique enables translation invariance at low computational cost compared to convolutions. This method also allows to reduce the dimension of the event stream without rescaling.

A similar method was already introduced in [12] for classifying a dataset of three human motions (bend, sit/stand, walk) recorded with a DVS. Their approach consists of remapping the address of their feature neurons (C1) with respect to their mean activation before being fed to the classifier. Instead, our method consists of remapping the address events directly, with respect to the median event. Unlike the median, the mean activation can result in an event-less attention window in case of multiple objects in motion, such as two-hand gestures. Additionally, since our attention window is smaller than the event stream, eccentric events are not processed by the network. We show in this paper how this biologically motivated technique boosts the performance, even on DvsGesture, where multiple body parts are simultaneously in motion. We note that a similar mechanism could be integrated in a robotic head as the one used in this paper to perform actual eye movements (see Figure 2). However, an additional mechanism to discard events resulting of the ego-motion would be required, which could be based on visual prediction [11], [13], [14].

D. Microsaccadic eye-movements

For our real-world grasping experiment, address events are sensed from static scenes by performing microsaccadic eye movements. This technique was already used to convert images to event streams [15], essentially extracting edge features [16]. To this end, we mounted the DVS on the robotic head presented in [16], see Figure 2. One Dynamixel servo MX-64AT is used to tilt both DVS simultaneously, while two other Dynamixel servos MX-28AT are used to pan each DVS independently. The center of all rotations is approximately the optical center of each DVS. In this work, only the events of the right DVS are processed. The microsaccadic motion consists of an isosceles triangle in joint space. The three motions defined by the edges of the triangle last 0.2 s each. The first motion consists of a negative tilt of α and negative pan of $\alpha/2$. The second motion is a tilt of α and negative pan of $\alpha/2$. The third motion moves the DVS back to its initial position with a pan of α. We chose the angle $\alpha = 1.833°$. This angle is much smaller in biology, but DVS pixels are much larger than the photoreceptors of the retina [17].

The microsaccades are triggered either manually for recording training data, or automatically in a loop at test time. We allow the events to flow through the network only when a microsaccade is triggered. No information about the properties of the microsaccade is passed as an input to the network.

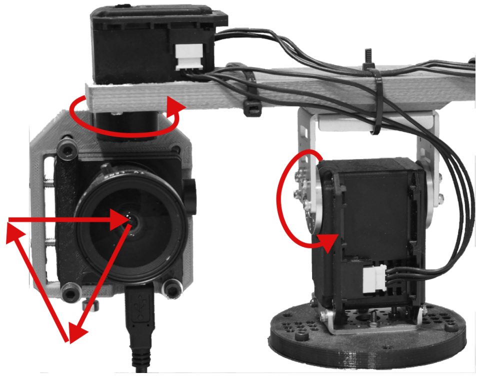


Figure 2: Microsaccadic motion of the DVS performed by the robotic head. The motion consists of three phases. An negative tilt of α and negative pan of $\alpha/2$, followed by a tilt of α and negative pan of $\alpha/2$, followed by a final pan of α moving the DVS back to its initial position. We chose the angle $\alpha = 1.833°$. Each motion is effectuated in 0.2 s.

III. Evaluation

Throughout the experiments, we rely on a dense 4-layers architecture with two hidden layers of 200 neurons respectively (see Figure 1). As described in Section II-B, the ON- and OFF-events obtained from the DVS are segregated in two separate input layers. All synapses are stochastic, with a 35% chance of dropping each spike. All the presented experiments relied on the open-source implementation of eRBP[1] based on the neural simulator Auryn [18].

A. DvsGesture

DvsGesture is an action recognition dataset recorded by IBM using a DVS [8], [19]. We reduce the dimensionality of the input event stream to 64×64, both in the rescaling case and the covert attention window case. It consists of 29 subjects performing 11 diverse actions in three different illumination conditions. We split the whole dataset into 1176 training samples and 288 test samples, leaving out the data when users do not perform a labeled motion. This training set consists of 7602s (approximately 2h) of recordings, versus 1960s for the test set. The duration of the actions varies greatly across samples, see Table I and Figure 3. A single sample may be about 1 second or over 18 seconds long. Despite this high variance, all samples were used in full without temporal modifications, both for training and testing. The number of events to calculate the position of the attention window was set to $n_{\text{attention}} = 1000$.

Our evaluation on DvsGesture shows that eRBP efficiently learns to classify motions from raw event streams with the attention mechanism introduced in Section II-C. The accuracy

[1] https://gitlab.com/eneftci/erbp_auryn

Label	#Training	#Test	Description
1	97	24	Clapping
2	98	24	Right hand waving
3	98	24	Left hand waving
4	98	24	Right arm circling clockwise
5	98	24	Right arm circling anticlockwise
6	98	24	Left arm circling clockwise
7	99	24	Left arm circling anticlockwise
8	196	48	Arms rolling
9	98	24	Air drum
10	98	24	Air guitar
11	98	24	Other gestures

Table I: Labels of all the different classes in the DvsGesture dataset and their description. The amount of samples of label 8 is doubled, since arms rolling were recorded and labeled with 8 for both rotation directions.

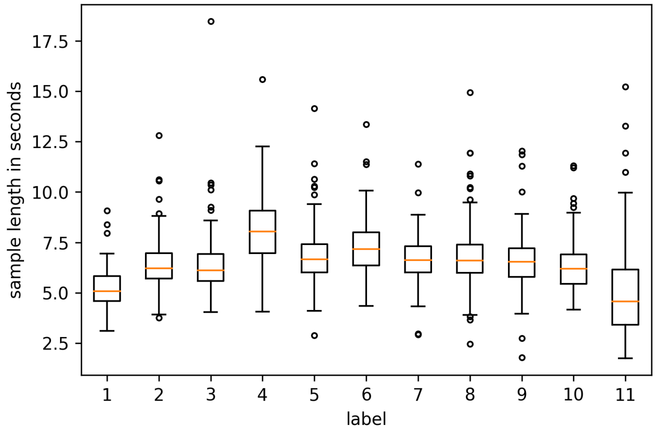
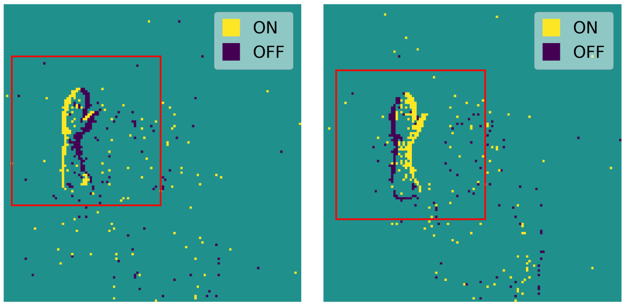


Figure 3: Statistics on sample duration for each label in the DvsGesture dataset. The red horizontal lines represent the median sample duration. Each box indicates the interquartile range ($IQR = Q_3 - Q_1$) per label, which is the range between the first Q_1 and the third quartile Q_3. Whisker pairs show the range of all sample durations within $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$. Outliers are represented by small circles.



(a) Arm circling clockwise (b) Arm circling anticlockwise

Figure 4: Aggregation of 1000 events for two samples of the DvsGesture dataset (user 10). Yellow pixels symbolize ON-events, blue pixels are OFF-events. The information about direction of motion is contained in the event polarity, hence the importance of their segregation in the input layer. The red square represents the attention window of size 64×64, calculated as the median of the last 1000 events.

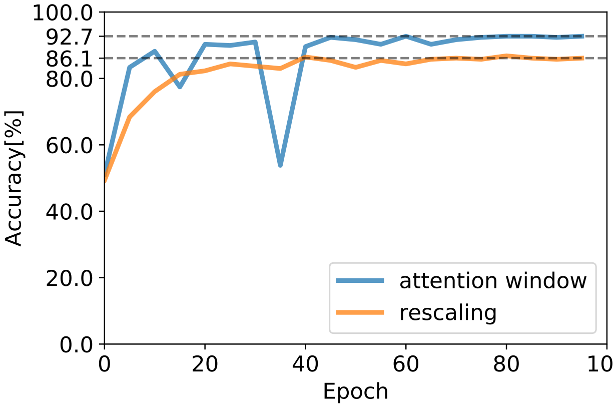


Figure 5: Classification accuracy over 100 epochs of training on DvsGesture. The accuracy converge to 86.1% and 92.7% for the rescaling and the attention window approaches respectively. The dimension of the event stream is 64×64 in both cases. The drop in accuracy in early stages of learning in the attention window case are due to the stochastic synapses, which have 35% chances of dropping spikes. The rescaling approach is resilient to this stochasticity since all events in the original stream are squeezed into macro-pixels, leading to redundant events.

of 92.7% is achieved after only 60 epochs, corresponding to approximately 127h of training data, see Figure 5. This accuracy is comparable to state-of-the-art deep networks (IBM EEDN [8], 94.49%) and other synaptic learning rules taking temporal dynamics into account (DCLL [3], 94.19%), both relying on convolutions. Without the attention mechanism, the accuracy of the network drops to 86.1%. These results confirm that our simple covert attention mechanism provides translation invariance without convolutions, at a low computational cost. Additionally, unambiguous samples are classified in under 0.1 s, with an increasing confidence over time, see Figure 6.

Since DvsGesture is a classification task, not accounting for neural temporal dynamics in the learning rule (Equation (1)) does not impact the performance much. Indeed, the target output signal as encoded by label neurons is constant across a training sample for durations of several seconds, see Figure 3. We expect this omission to decrease performance significantly for a temporal regression task – such as learning a time sequence – where the temporal dynamics of the target signal is relevant.

B. *Grasp-type Recognition*

In this experiment, we embody eRBP in the real-world grasping robotic setup depicted in Figure 7. In this setup, the spiking network is trained to recognize four labels corresponding to four different types of grasp: ball-grasp, bottle-grasp, pen-grasp or do nothing [20]. During training, an object of a particular class is placed on a table at a specific position. The robotic head performs microsaccadic eye movements (similar to the N-MNIST dataset [15]) to extract visual information

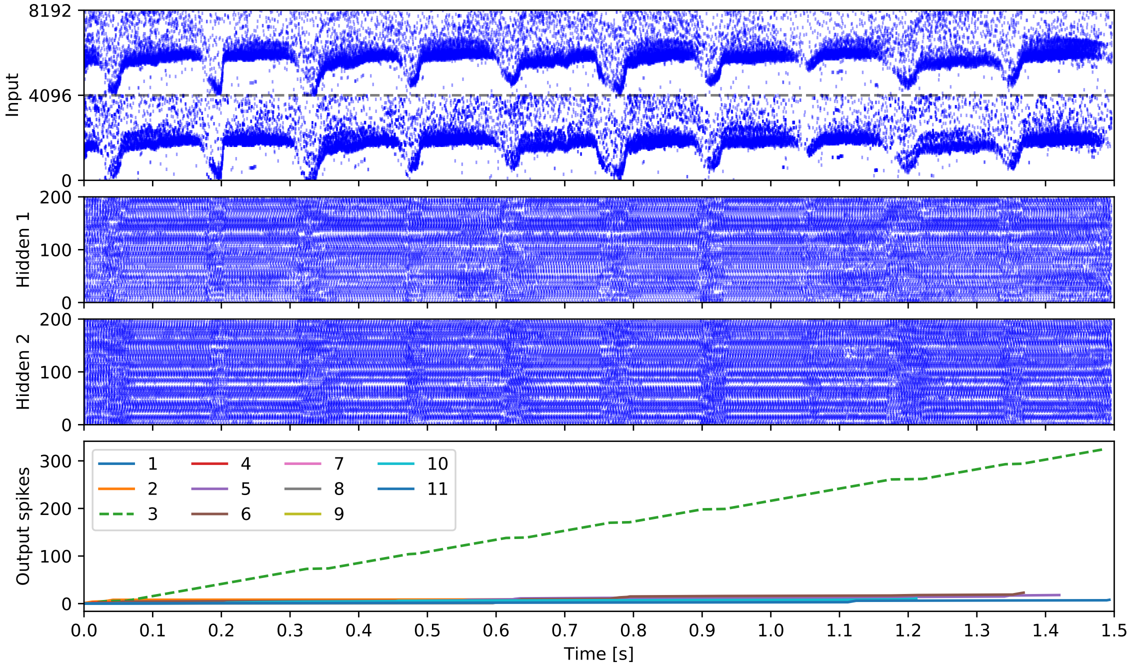


Figure 6: Spiketrains and classification results for a test sample of class “left hand waving” from DvsGesture dataset. The network manages to correctly classify the sample in less than $0.1\,\mathrm{s}$, with increasing confidence over time. The rhythm of the “left hand waving” motion is clearly visible in the input spiketrain. The neurons in the hidden layers spike close to their maximum frequency, as limited by the refractory period.

from the static object. The event streams are recorded together with the corresponding object affordance. In this experiment, the attention window has dimension 32 × 32 and is fixed to match the position of the objects on the table, see Figure 8 for example samples. During testing, a microsaccade is performed and the detected object affordance triggers the adequate predefined reaching and grasping motion on a Schunk LWA4P arm equipped with a five-finger Schunk SVH gripper. This demonstrator was implemented with the ROS Framework [21] and the ROS DVS driver introduced in [22].

With only 20 samples per class and 30 training epochs, the network was capable of learning the four visual affordances, see the supplementary video[2]. Example spiketrains and classification results at test time are shown in Figure 9. The network classifies the ball and the pen with high confidence, but with low confidence for the bottle and the background. This is due to the fact that we trained with a transparent bottle, generating only a few events, thus visually similar to the background.

While the microsaccade leads to sparse activity in the input layer, both hidden layers have strong, sustained activity (see Figure 9). This indicates some form of short-term memory emerging from neural dynamics, as the network is capable of remembering the object despite receiving no events between two microsaccadic phases. The neurons in the hidden layers spike close to their maximum frequency, as limited by the refractory period. This is not surprising, since we did not use a regularization term to enforce sparsity in the learning rule and labels were encoded into spikes with rate coding, see Section II-B. Except for the background class, counting the spikes at the output leads to successful classification about 100ms after microsaccade onset, comparable to human performance reported in behavioral study [23].

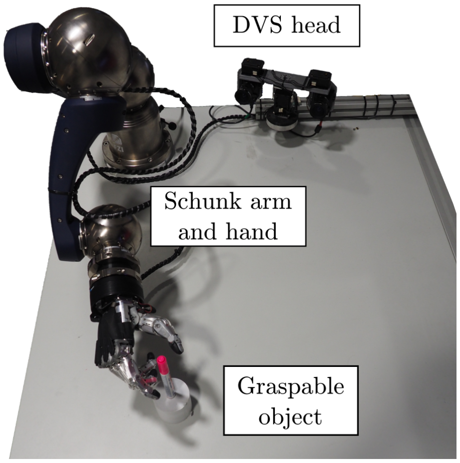


Figure 7: Real-world grasp-type Recognition experiment setup integrating a Schunk LWA4P arm equipped with a five-finger Schunk SVH gripper and a DVS head. The DVS head performs microsaccadic eye movements to sense event streams from static scenes. We recorded a small four-classes dataset (ball, bottle, pen, background) of 20 samples per class. At test time, the detected grasp-type triggers the corresponding predefined reaching and grasping motion.

[2] https://neurorobotics-files.net/index.php/s/sBQzWFrBPoH9Dx7

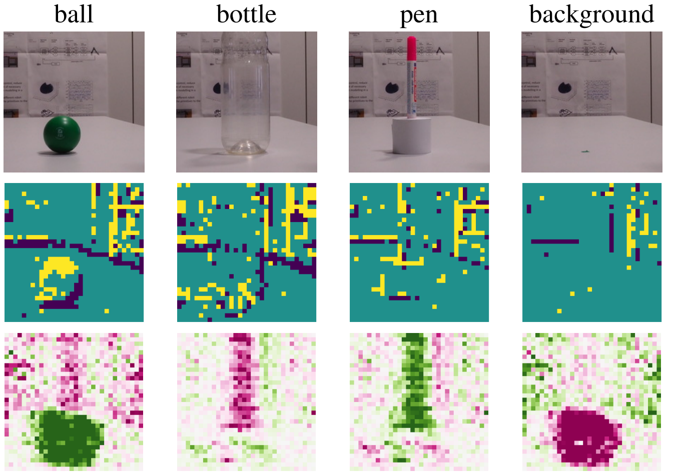


Figure 8: Example samples and learned weights for the grasp-type recognition experiment. Top row: camera image of the objects. Middle row: integration of the address events during 15ms after microsaccade onset. Bottom row: projection of the synaptic weights of our 4-layers network for each label neuron onto the input after training. Green denotes excitation (positive influence) and pink denotes inhibition (negative influence).

Since the DVS does not sense colors, the network only relies on shape information, crucial for affordances. This allowed the network to moderately generalize despite the small amount of training samples. Indeed, a single object per affordance was used during training, but the network could recognize different objects of the same shape. Recognition also worked when the objects were slightly moved from the reference point used for grasping. However, the network was not robust to change in background or unexpected background motions happening during the microsaccade. This is due to the background being learned as an additional class for the "do nothing" affordance.

IV. Conclusion

Neuromorphic engineering technology enables the design of autonomous learning robots operating at high speed for a fraction of the energy consumption of current solutions. Until recently, the advantages of this technology were limited due to the lack of efficient learning rules for spiking networks. This bottleneck has been addressed since the realization that gradient backpropagation can be approximated with spikes. In this paper, we demonstrated the ability of eRBP to learn from real-world event streams provided by a DVS. First, with the addition of a simple covert attention mechanism, we have shown that eRBP achieved comparable accuracy as state-of-the-art methods on the DvsGesture benchmark [8]. This attention mechanism improved performance by a large margin in comparison to classical rescaling approaches, by providing translation invariance at a low computational cost compared to convolutions. Second, we integrated eRBP in a real-world robotic grasping experiment, where affordances are detected from microsaccadic eye movements and conveyed to a robotic arm and hand setup for execution. Our results show that correct affordances are detected within about 100ms after microsaccade onset, matching biological reaction time [23].

For future work, eRBP could be replaced by more recent learning rules accounting for neural temporal dynamics [1]–[4]. This would enable the setup to be extended to temporal sequence learning and reinforcement learning tasks [24], [25]. Additionally, other components of the grasp-type recognition experiment could be implemented with spiking networks, such as reaching motions [26], [27], grasping motions [28] and depth perception [16]. This work paves the way towards the integration of brain-inspired computational paradigms into the field of robotics.

Acknowledgment

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).

References

[1] E. O. Neftci, C. Augustine, S. Paul, and G. Detorakis, "Event-driven random back-propagation: Enabling neuromorphic deep learning machines," *Frontiers in neuroscience*, vol. 11, p. 324, 2017.

[2] F. Zenke and S. Ganguli, "Superspike: Supervised learning in multi-layer spiking neural networks," *arXiv preprint arXiv:1705.11146*, 2017.

[3] J. Kaiser, H. Mostafa, and E. Neftci, "Synaptic plasticity dynamics for deep continuous local learning," *arXiv preprint arXiv:1811.10766*, 2018.

[4] G. Bellec, F. Scherr, E. Hajek, D. Salaj, R. Legenstein, and W. Maass, "Biologically inspired alternatives to backpropagation through time for learning in recurrent neural nets," *arXiv preprint arXiv:1901.09049*, 2019.

[5] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman, "Random synaptic feedback weights support error backpropagation for deep learning," *Nature communications*, vol. 7, p. 13276, 2016.

[6] M. Jaderberg, W. M. Czarnecki, S. Osindero, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Decoupled neural interfaces using synthetic gradients," *arXiv preprint arXiv:1608.05343*, 2016.

[7] Z. Bing, C. Meschede, F. Röhrbein, K. Huang, and A. C. Knoll, "A survey of robotics control based on learning-inspired spiking neural networks," *Frontiers in neurorobotics*, vol. 12, p. 35, 2018.

[8] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza *et al.*, "A low power, fully event-based gesture recognition system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7243–7252.

[9] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks," *arXiv preprint arXiv:1901.09948*, 2019.

[10] M. Zirnsak, N. A. Steinmetz, B. Noudoost, K. Z. Xu, and T. Moore, "Visual space is compressed in prefrontal cortex before eye movements," *Nature*, vol. 507, no. 7493, p. 504, 2014.

[11] M. A. Sommer and R. H. Wurtz, "Influence of the thalamus on spatial visual processing in frontal cortex," *Nature*, vol. 444, no. 7117, p. 374, 2006.

[12] B. Zhao, R. Ding, S. Chen, B. Linares-Barranco, and H. Tang, "Feedforward Categorization on AER Motion Events Using Cortex-Like Features in a Spiking Neural Network." *IEEE transactions on neural networks and learning systems*, vol. PP, no. 99, p. 1, 2014.

[13] J. Kaiser, R. Stal, A. Subramoney, A. Roennau, and R. Dillmann, "Scaling up liquid state machines to predict over address events from dynamic vision sensors," *Bioinspiration & biomimetics*, vol. 12, no. 5, p. 055001, 2017.

[14] J. Kaiser, S. Melbaum, J. C. V. Tieck, A. Roennau, M. V. Butz, and R. Dillmann, "Learning to reproduce visually similar movements by minimizing event-based prediction error," in *Int. Conf. on Biomedical Robotics and Biomechatronics*. IEEE, 2018.

[15] G. Orchard, A. Jayawant, G. Cohen, and N. Thakor, "Converting static image datasets to spiking neuromorphic datasets using saccades," *arXiv preprint arXiv:1507.07629*, 2015.

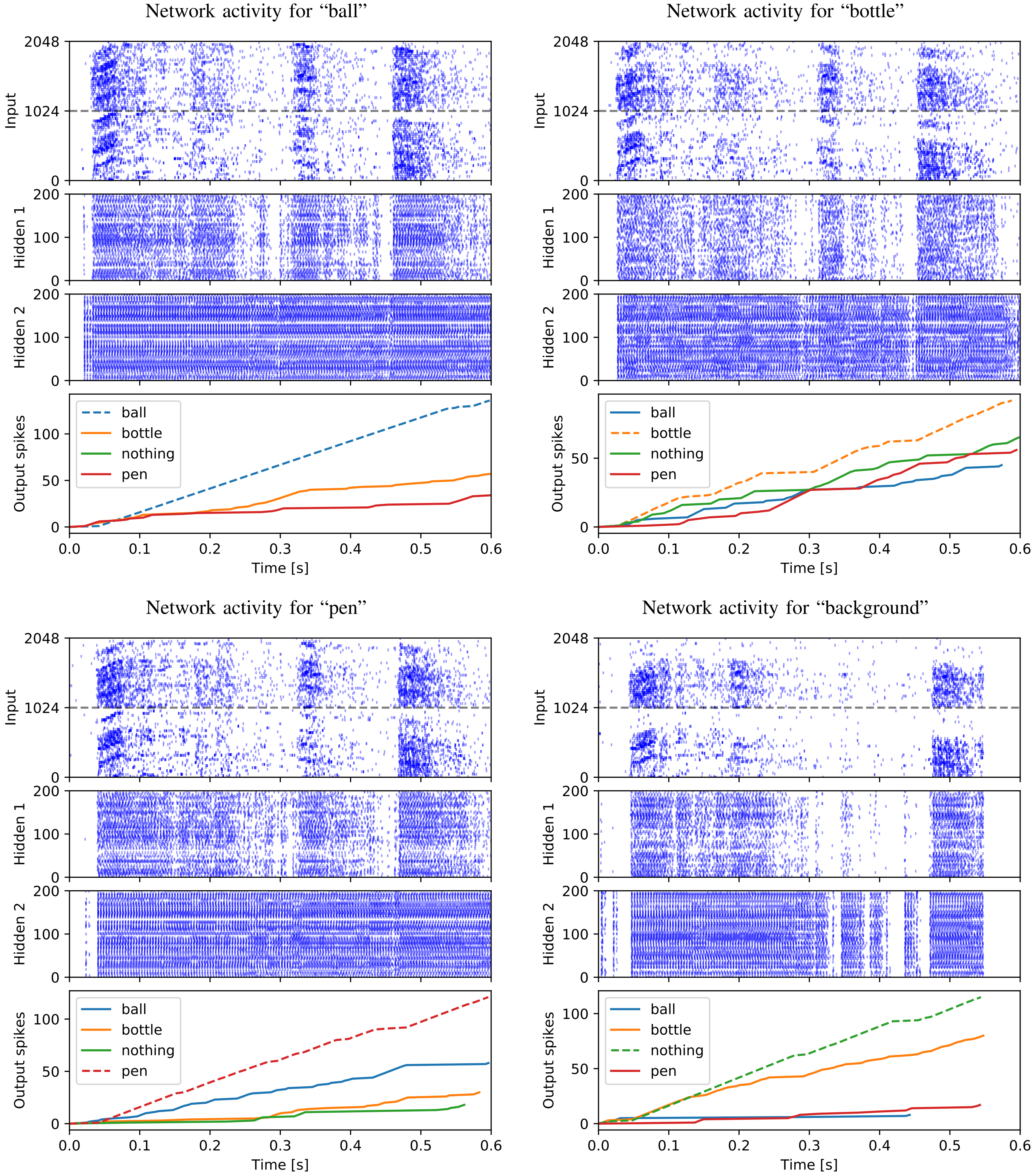


Figure 9: Spiketrains and classification results for four test samples of our grasp-type dataset. The network manages to correctly classify the four test samples. The ball and the pen are easily classified despite the small amount of training data. However, the transparent bottle generates few events, yielding to more uncertainty in the classification with the background. The three phases of the microsaccadic motion are clearly visible in the input spiketrains (first row of each plot), see Section II-D. However, the activity of the hidden layers does not drop instantaneously even when the input is sparse, indicating a form of short-term memory induced by the neural dynamics. The neurons in the hidden layers spike close to their maximum frequency, as limited by the refractory period.

[16] J. Kaiser, J. Weinland, P. Keller, L. Steffen, J. C. V. Tieck, D. Reichard, A. Roennau, J. Conradt, and R. Dillmann, “Microsaccades for neuromorphic stereo vision,” in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 244–252.

[17] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel, “The role of fixational eye movements in visual perception,” *Nature Reviews Neuroscience*, vol. 5, no. 3, pp. 229–240, 2004.

[18] F. Zenke and W. Gerstner, “Limits to high-speed simulations of spiking neural networks using general-purpose computers,” *Frontiers in neuroinformatics*, vol. 8, p. 76, 2014.

[19] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128x128 120 db 15us latency asynchronous temporal contrast vision sensor,” *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[20] J. Kaiser, D. Zimmerer, J. C. V. Tieck, S. Ulbrich, A. Roennau, and R. Dillmann, “Spiking convolutional deep belief networks,” in *International Conference on Artificial Neural Networks*. Springer, 2017, pp. 3–11.

[21] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, “Ros: an open-source robot operating system,” in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.

[22] E. Mueggler, B. Huber, and D. Scaramuzza, “Event-based, 6-dof pose tracking for high-speed maneuvers,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2761–2768.

[23] J. G. Martin, C. E. Davis, M. Riesenhuber, and S. J. Thorpe, “Zapping 500 faces in less than 100 seconds: Evidence for extremely fast and sustained continuous visual search,” *Scientific reports*, vol. 8, no. 1, p. 12482, 2018.

[24] J. C. V. Tieck, M. V. Pogančić, J. Kaiser, A. Roennau, M.-O. Gewaltig, and R. Dillmann, “Learning continuous muscle control for a multi-joint arm by extending proximal policy optimization with a liquid state machine,” in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 211–221.

[25] J. Kaiser, M. Hoff, A. Konle, J. C. V. Tieck, D. Kappel, D. Reichard, A. Subramoney, R. Legenstein, A. Roennau, W. Maass, and R. Dillmann, “Embodied synaptic plasticity with online reinforcement learning,” *Frontiers in neurorobotics*, p. Submitted, 2019.

[26] J. C. V. Tieck, L. Steffen, J. Kaiser, D. Reichard, A. Roennau, and R. Dillmann, “Combining motor primitives for perception driven target reaching with spiking neurons,” *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, vol. 13, no. 1, pp. 1–12, 2019.

[27] J. C. V. Tieck, L. Steffen, J. Kaiser, A. Roennau, and R. Dillmann, “Controlling a robot arm for target reaching without planning using spiking neurons,” in *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2018, pp. 111–116.

[28] J. C. V. Tieck, H. Donat, J. Kaiser, I. Peric, S. Ulbrich, A. Roennau, M. Zöllner, and R. Dillmann, “Towards grasping with spiking neural networks for anthropomorphic robot hands,” in *International Conference on Artificial Neural Networks*. Springer, 2017, pp. 43–51.