Neuromorphic foveation applied to semantic segmentation

Amélie Gruel[1] Dalia Hareb[1] Jean Martinet [1] Bernabé Linares-Barranco[2] Teresa Serrano-Gotarredona[2]

[1]Université Côte d'Azur, CNRS, I3S, France
[2]Instituto de Microelectrónica de Sevilla IMSE-CNM, Sevilla, Spain

Abstract

Foveation can be defined as the organic action of directing the gaze towards a visual region of interest, to selectively acquire relevant information. With the recent advent of event cameras, we believe that taking advantage of this visual neuroscience mechanism would greatly improve the efficiency of event-data processing. Indeed, applying foveation to event data would allow to comprehend the visual scene while significantly reducing the amount of raw data to handle. In this respect, we present in this work the evolution of the performance of semantic segmentation with respect to the amount of event data used, to demonstrate the stakes of foveation.

1. Introduction

Event cameras (or silicon retinas) represent a new kind of sensors that measure pixel-wise changes in brightness and output asynchronous events accordingly [8]. This novel technology allows for an energy-efficient recording and storage of data evolving over time and space. Indeed each event is recorded punctually and asynchronously with no redundancy; as opposed to traditional frame-based cameras, where each pixel outputs data in all frames, in a synchronous manner.

Spiking neural networks (SNN) are artificial neural networks which mimic the dynamics of biological neuronal circuits by receiving and processing information in the form of spike trains. They are particularly well suited to handle the atypical kind of data output from event cameras, since each event can be assimilated to an activation spike between two spiking neurons.

Foveation is the biological action allowed by the structure of the eye [2], driven by the visual attention [4]. When the gaze is directed towards a region of interest (RoI), the center of the perceptual field is caught by the fovea, a small and central spot in the retina where the vision is optimal in bright light. The further away we get from the fovea, the lesser information is processed by the eye.

We believe that developing a mechanism approaching foveation would greatly improve event data processing, going beyond its significant energy-efficiency. Indeed, this would allow to maintain a high accuracy regarding relevant information, while significantly reducing the amount of raw data to handle. Since the energy consumption of neuromorphic architectures such as SpiNNaker or Intel Loihi is directly linked to the number of spikes/events processed, reducing this number is an efficient way of reducing the consumption, which is central for embedded applications. Furthermore, this approach consistency is supported by the fact that silicon retinas aim by definition to reproduce the biological retina behaviour.

To demonstrate the interest of applying foveation to event data, we propose to study the respective evolutions of the amount of event data processed by a segmentation algorithm and its accuracy when foveation is applied. In order to simulate the foveation, the event data will be processed at a higher or lower resolution, depending on the relevance of the spatial regions in the image at different coordinates. Our proposed model goes beyond biology by allowing multiple RoI of arbitrary size and shape. This approach is part of the work conducted in the context of APROVIS3D[1] project, and is ultimately to be applied to the use case of coastline tracking by a UAV. Thus the dataset is chosen in order to approach this use case, as well as the segmentation task.

2. Foveation methodology

2.1. Saliency detection

The detection of RoI to foveate on is a little-explored issue regarding event-data. In this work, we propose to use

[1]URL: http://aprovis3d.eu

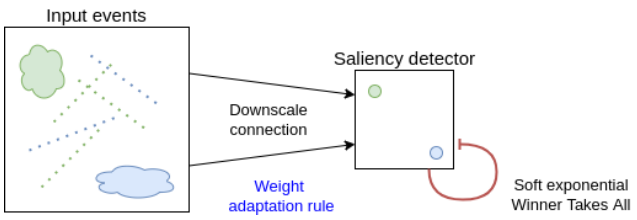


Figure 1. Spiking neural network model used to detect saliency by event density, adapted from [6].

part of the SNN presented in [6]: this saliency detector integrates the events produced by each pixel at a low resolution and outputs a set of coordinates for one or multiple RoI. In this case, our RoI would be a region where the amount of events received over a certain amount of time is more important than elsewhere over the whole scene.

This whole mechanism relies solely on intrinsic SNN dynamics and dynamic adaptation rules applied to synaptic weights and population thresholds. This is a crucial feature as it leads to minimising the latency since it does not require the conversion of spiking events into a frame. The saliency detection is not specialised for any specific context or any specific shape, which allows for a good generalisability of the network. The proposed architecture, shown in Fig. 1, is designed to be lightweight enough to enable running in real-time.

Input layer The input layer translates sensor relative changes in the illumination (or events) into spikes. The spikes produced by the input layer are sent to the saliency detector via an excitatory downscaling connection. This corresponds to a convolutional layer with a kernel size $S \times S$, a stride S, without padding. The input neurons are separated into non-overlapping square regions of size $S \times S$. Each neuron in the input layer's subregions is connected to one corresponding neuron in the saliency detector layer.

Saliency detector The saliency detection aggregates the active regions into distinct segments using a soft Winner-Takes-All (WTA) by laterally inhibiting the neurons in the same layer: each neuron activation leads to the inhibition of the others, without autapses (self-connections). Since a strong WTA leads to the activation of only one neuron in the layer and multiple RoI are to be detected by the network, the soft WTA weight has been set experimentally to 0.02.

In the case of the saliency detector, a specific exponential WTA is implemented according to the radial basis function Eq. 1 in order to allow RoI of arbitrary sizes:

$$W_{WTA} = max(\frac{e^d}{w \times h}, w_{max}) \tag{1}$$

where d corresponds to the Euclidean distance in number of neurons between the active and target neuron subject to inhibition, and w and h to the width and height of the layer. The weight W_{WTA} has an upper bound of $w_{max} = 50$.

Finally, the adaptive detection of saliency in this layer is enabled by a dynamic weight adaptation rule between the input layer and the saliency detector, inspired by Hebb's rule: "cells that fire together wire together" [7]. This rule is implemented by increasing or decreasing the weights of synapses that have recently fired.

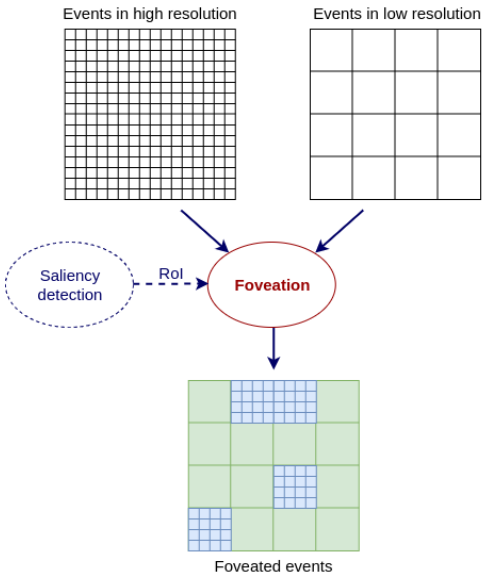


Figure 2. Binary foveation of events using the corresponding high and low resolution (spatially downscaled by $factor$) and based on a known region of interest, delimited by the MIN and MAX points.

2.2. Reconstitution of foveated data

In this work, we consider the foveation process akin to the combination of a sample's events in high resolution and low resolution using a mask, as presented by the Fig. 2. This binary combination discriminates the fovea (events in high resolution – blue region in Fig. 2) from the retinal periphery (low resolution; i.e. spatially downscaled – green region in Fig. 2). The RoI (in red in Fig. 2) detected by the saliency detector mentioned earlier is thus assimilated to the fovea.

Let (x_{min}, y_{min}) and (x_{max}, y_{max}) be the coordinates of the delimiting points of the area of foveation detected by the saliency detector (as seen on Fig. 2),

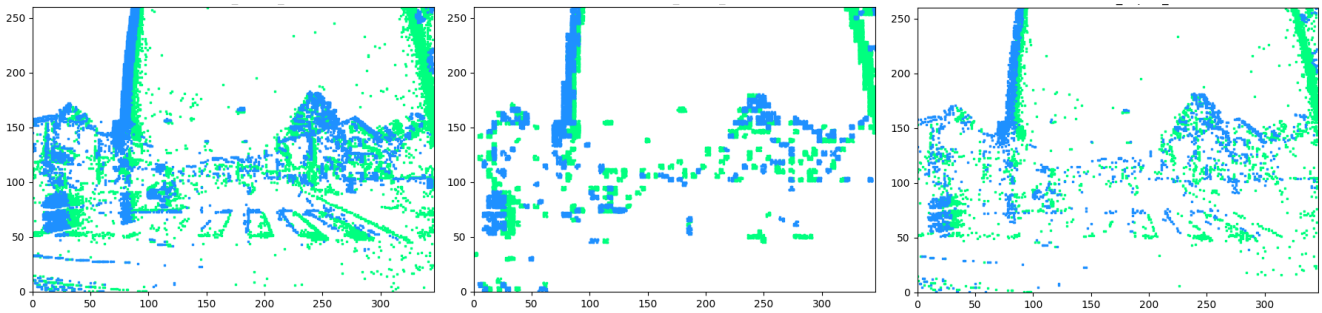
$$\begin{aligned} Fovea &= \{(x,y) | x \in [x_{min}, x_{max}], y \in [(y_{min}, y_{max}]\} \\ Periphery &= \{(x,y) | x \notin [x_{min}, x_{max}], y \notin [(y_{min}, y_{max}]\} \end{aligned} \tag{2}$$

where $Fovea$ and $Periphery$ correspond to the coordinates of the set of salient and non-salient events respectively, in different resolutions.

2.3. Event data reduction

Event data reduction is not trivial, as explained in [5]. Many different approaches can be used to produce the spatial downscaling depicted in Fig. 2. We decided to use the log-luminance reconstruction using event count method described in [5]. In a nutshell, it relies on assimilating event spatial downscaling to averaging the luminance captured by a subset of pixels in the original sensor.

(a) Events in the sample in high resolution (left), low resolution (middle) and after foveation (right).



(b) Segments and performance values identified by Ev-SegNet in the sample in high resolution (left), low resolution (middle) and after foveation (right).

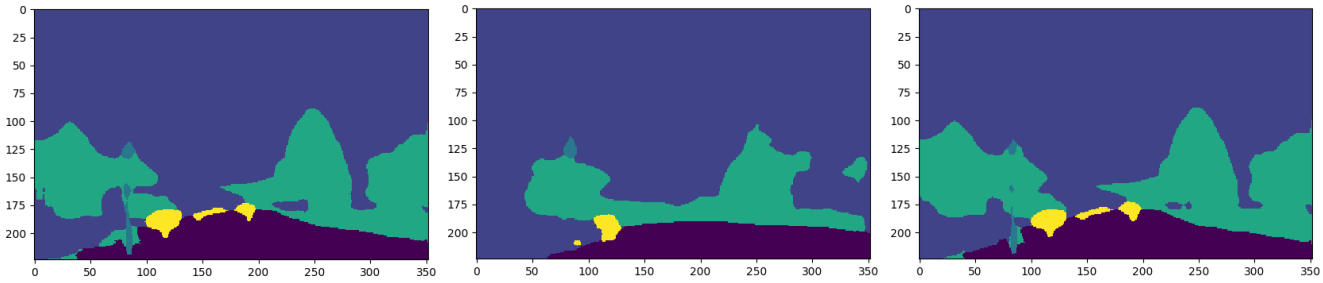


Figure 3. **Top.** Visual representation of the events in the sample `rec1487417411_export_1467` after various processes. **Bottom.** Visual representation of the different segments identified by Ev-SegNet in the same sample.

It is to be noted that in order to process high and low resolution events using the same frame of reference, an expansion was applied to the spatially reduced data so that a reduced pixel physically corresponds to the size of $factor \times factor$ original pixels.

3. Experimental validation

To validate our proposed model, we select the task of semantic segmentation, which is is a computer vision task in which specific regions of an image are labelled according to its semantic contents, a key task for scene understanding. It has been extensively studied using artificial neural networks, more specifically, Convolutional Neural Network (CNN) model with either frames or events as input.

This section describes the dataset and the semantic segmentation model used to perform such an experiment, as well as the comparative results.

3.1. Event-based dataset

The DAVIS Driving Dataset 2017 (DDD17) [3] contains 40 different driving sequences of event data captured by an event camera. However, since the original dataset provides only both grayscale images and event data without semantic segmentation labels, we used the segmentation labels provided in [1] that uses 20 different sequence intervals taken from 6 of the original DDD17 sequences. Furthermore, as only multi-channel representation of the events (normalized sum, mean and standard deviation for each polarity) are made available, we extracted the original events from DDD17 with the traditional $< x, y, p, t >$ structure using DDD20 tools[2] and selected the ones corresponding to the frames that have a ground truth. The resulting dataset is split into a training dataset consisting of 15,950 frames and a testing one consisting of 3,890 frames.

As presented in the Fig. 3a, the event data's properties were compared for sample in high resolution (original dataset), low resolution (spatially downscaled with factor 4) and foveated (binary combination of the previous two).

3.2. Semantic segmentation model

In our work, we used the model built by [1] for outperforming all existing studies in this kind of task using event cameras. This model is inspired from current state-of-the-art semantic segmentation CNNs, slightly adapted to use the event data encoding. It consists of an encoder-decoder architecture: an encoder represented by Xception model in which all the training is concentrated, and a light decoder connected to the encoder via skip connections to help deep neural architecture to avoid the vanishing gradient problem. The use of an auxiliary loss increases convergence speed.

The model takes as input 6 channels representing the

[2] https://github.com/SensorsINI/ddd20-utils

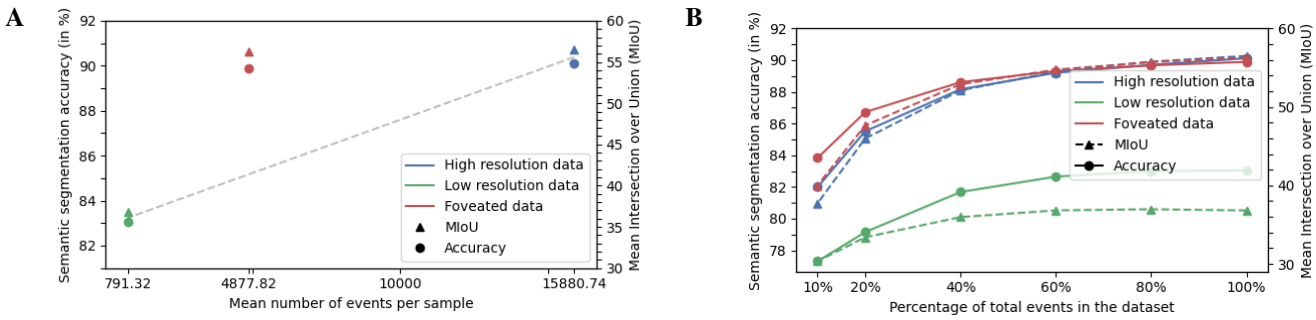


Figure 4. Semantic segmentation accuracy according to the number of events in the dataset after processing (left) and its evolution according to the percentage of total events in the dataset after processing (right) for the event data in high resolution (in blue), low resolution (in green) and after foveation (in red).

count, mean and standard deviation of the normalized timestamps of events happening at each pixel, included in the selected frames described in the section 3.1 within an interval of $50ms$ for the positive and negative polarities. Finally, the training is performed via backpropagation in order to minimise the soft-max cross-entropy loss measured by summing the error between the estimated pixels' classes and the true ones.

3.3. Results

Figure 4 presents a comparison between the different versions of the dataset, i.e. the DDD17 dataset in high and low resolutions and after foveation, according to the accuracy and the MIoU (Mean Intersection over Union) of the semantic segmentation performed by Ev-SegNet, which equations are described in [1].

To validate our initial hypothesis, the foveation would have to produce a number of events significantly closer to the low resolution's while allowing for a semantic segmentation performance closer to the high resolution's. In other terms, the foveated results should be above the dotted grey line on Fig. 4a. We do observe a striking decrease in the number of events between pre- (high resolution) and post-processing (low resolution and foveation) of the dataset; the spatial downscaling keeps 5% of the original events while the foveation includes 30% of events. Similarly, the foveation's accuracy and MIoU are remarkably close to the high resolution's performance. Those two observations combined do validate our core thesis.

Furthermore, it is interesting to note that when comparing the proportional decrease of the number of events in the dataset post-process in Fig.4b, while all three types of data show the same behaviour, the foveated data outperforms the high resolution data from an 80% decrease and downwards. This is explained by the fact that the majority of events kept in the foveated dataset provides relevant information to the semantic segmentation model, while a significant part of the events in the original dataset is not as useful.

4. Conclusion

In this work, we demonstrate the stakes of foveation applied to event data for semantic segmentation. Such a strategy does concurrently preserve the accuracy of event data processing and greatly reduce the amount of data needed for the task. Further research will validate the proposed approach with several levels of resolution – not only 2, and for other tasks e.g. classification.

Acknowledgements

This work was supported by the European Union's ERA-NET CHIST-ERA 2018 research and innovation programme under grant agreement ANR-19-CHR3-0008.

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

[1] I. Alonso and A.C. Murillo. EV-SegNet: Semantic Segmentation for Event-based Cameras. *CVPR W.*, 2019. 3, 4

[2] M.F. Bear et al. The Human Eye. In *Neurosciences, Exploring the brain*, Wolters Kluwer Health. 2007. 1

[3] J. Binas, D. Neil, S.C. Liu, and T. Delbruck. Ddd17: End-to-end davis driving dataset. *arXiV*, 2017. 3

[4] A. Gruel and J. Martinet. Bio-inspired visual attention for silicon retinas based on spiking neural networks applied to pattern classification. *CBMI*, 2021. 1

[5] A. Gruel, J. Martinet, T. Serrano-Gotarredona, and B. Linares-Barranco. Event data downscaling for embedded computer vision. In *VISAPP*, 2022. 2

[6] A. Gruel, A. Vitale, J. Martinet, and M. Magno. Neuromorphic event-based spatio-temporal attention using adaptive mechanisms. In *AICAS*, 2022. 1, 2

[7] D.O. Hebb. The organization of behavior: A neuropsychological theory. *Journal of the American Medical Association*, 143(12), 1949. 2

[8] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128x128 120 db 15 ms latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2), 2008. 1