EE367 - Project Proposal

A SIREN-based framework for events to video reconstruction

Axel Levy

1 Motivation and goal of the project

A dynamic vision sensor (DVS), also known as event camera or neuromorphic camera, is an imaging sensor that responds to local changes in irradiance (brightness). In a DVS, each pixel operates independently and asynchronously by storing a reference level of irradiance and continuously comparing it to the current level of irradiance. If the difference in irradiance on a pixel exceeds a given threshold, that pixel resets its reference level and generates an *event* (Fig. 1). The output of a DVS is therefore a stream of events, where each event is characterized by the *coordinates* of the pixel that generated the event, the *time* when it was generated and its *polarity* (positive if the irradiance was increasing, negative otherwise). The main advantages of this type of sensor is its high dynamic range (up to 120 dB) and its high temporal resolution (each pixel can fire an event every microsecond). However, these advantages can only be exploited if a video (series of images) can be recovered from a stream of events.

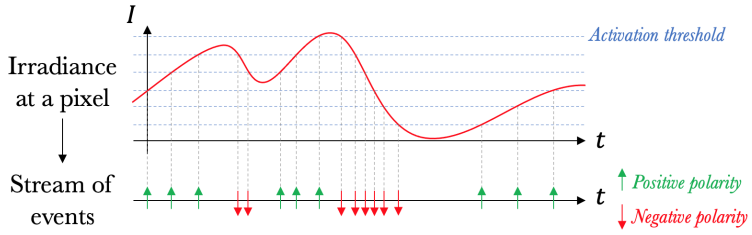

Figure 1: Working principle of a DVS. At each pixel, the temporal signal of irradiance is converted into a stream of events.

.

The purpose of this project is to propose a new framework to convert the output of a DVS into a conventional video stream. In this framework, an Implicit Representation Network is used to represent the data we want to recover (the video stream). Doing so, based on a quasi-continuous signal, we build a continuous representation of a video.

2 Related Work

The most simple way to recover a video stream is to integrate, at each pixel, the stream of (signed) events [1]. This technique is straightforward to implement, but only works if there is a point in time where we know the irradiance received by each pixel. A more advanced technique estimates the spatial gradients of images using a stream of events and recovers a series of images by Poisson integration [2], [3]. Although this technique is more robust to pixel mismatch (difference in threshold levels across pixels), it requires to batch the events into bins of fixed temporal width before performing Poisson integration. Therefore, the temporal resolution of the video is fixed. Finally, the *Contrast Maximization* technique [4] proposes to estimate motion, depth and optical flow by maximizing the contrast of an image of warped events. Although this technique can be applied to a diverse set of important vision tasks with event cameras, it does not directly perform video reconstruction, and an additional step (Poisson integration with temporal binning) is required to perform this task.

Unlike previous methods, some techniques do not rely on any prior knowledge about how an event camera works. This is for example the case with the *E2VID* framework [5] where a recurrent network converts a stream of events into a conventional video stream. This network needs to be trained on a large amount of simulated event data. Moreover, events are batched into temporal bins before being fed into the network, which means that part of the temporal resolution captured by the DVS is lost.

The framework we propose uses a SIREN to continuously represent the video stream. The implicit representation network is trained by comparing its output with the output of the DVS. This implies that we must design a relevant loss function to compare a video stream and a stream of events. Unlike previously described methods, a SIREN-based framework offers a temporally continuous representation of the video and completely preserves the temporal resolution captured by the DVS. By comparing the output of the SIREN and the output of the DVS, this framework is necessarily inspired by the forward model of the DVS. Finally, the framework we propose is highly flexible and can easily be modified to take into account additional inputs like conventional images.

3 Overview of the Framework

In a stream of events $\mathcal{E} = \{e_i\}_i$, each event e_i is characterized by its spatio-temporal coordinates (x_i, y_i, t_i) and its polarity $p_i \in \{-1, 1\}$. The framework is schematized in Fig. 2: the SIREN takes spatio-temporal inputs and represents the video stream. The video stream is compared with the stream of events *via* the loss function $\mathcal{L}$. The loss is backpropagated to train the network.

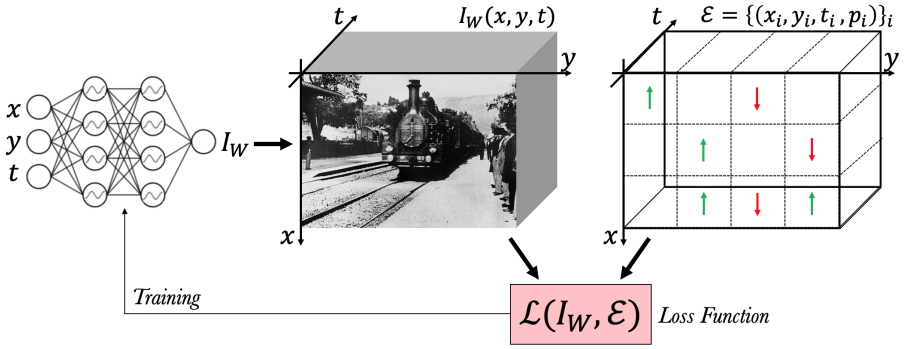

Figure 2: SIREN-based framework.

.

We will try using different loss functions to train the network. For a given pixel (x, y) and a given time t, one idea could be to sum the events that are fired between t and $t+\Delta t$. This sum $s(x, y, t)$ multiplied by the difference between two thresholds (δI) should match the difference in irradiance between $I(x, y, t+\Delta t)$ and $I(x, y, t)$ (with an uncertainty of $2\delta I$). An example of loss function could therefore be:

$$\mathcal{L}(x, y, t; I_W, \mathcal{E}) = (I(x, y, t + \Delta t) - I(x, y, t) - \delta I \cdot s(x, y, t))^2 + \text{regularization} \tag{1}$$

4 Milestones and Timeline

- *Already Done*: Preliminary results with gradient-based and finite difference methods (anisotropic TV regularization). Comparative discrete baseline (anisotropic and isotropic TV regularization). Theoretical study of the finite difference method.
- *Week 1*: Record new data with a DVS (DAVIS). Finalize the comparative baseline and plot "quality" versus "memory". Tests with slightly different loss function (*e.g.* double hinge loss, vary bin sizes).
- *Week 2*: Try to regularize with perceptual loss and isotropic TV. Test various activation functions. Definitive choice of loss function, regularizer and activation function.
- *Week 3*: Wide literature review. Precise tuning of hyper-parameters to get visually best results (pre-factor for regularization, bin sizes, number of hidden features and hidden layers). Generate high-quality videos.
- *Week 4*: Generate high-quality videos. Write report and poster. Record the video.

References

[1] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018.

[2] Julien NP Martel. *Unconventional Processing with Unconventional Visual Sensing: Parallel, Distributed and Event Based Vision Algorithms & Systems*. PhD thesis, ETH Zurich, 2019. Chapters 5 & 6.

[3] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pages 770–776. IEEE, 2011.

[4] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2018.

[5] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 2019.