Denosing Event Data with Neural Networks

Tianshu Kuai, Yan Ma, Yihan Ni

November 2022

1 Motivation

Event cameras are bio-inspired sensors that mimic the working principle of human retinas. Unlike traditional cameras with rolling shutters that capture image frames at regular frequencies, event cameras detect pixel-wise binary brightness changes in the scene and output asynchronous sequences of "events". The advantages of event cameras compare to RGB cameras include a high dynamic range, microsecond-level temporal resolution, and no motion blur. Therefore, event cameras are well-suited for high-speed applications such as driving scenarios.

There are plenty of methods to denoise image data sampled from RGB cameras. However, they cannot be directly applied to event data due to the difference in data representations. In event data learning, deep-learning based approaches usually convert event streams to image-like data and use image models like CNNs to do further processing. We are interested in exploring deep neural networks' capabilities in terms of denoising event data. In addition, we will explore and compare the effectiveness of different event data representations that are commonly used in the community.

2 Related Works

2.1 Event Voxel Representation

Since the event data samples are streams of events, existing methods usually transform the raw event streams into image-like data that can be processed by CNNs. EST [4] proposed a way to convert the event data from a four-dimensional (x, y, t, p) structure, to a three-dimensional voxel grid by projecting or summing one of the four dimensions. It shows that the Event Voxel Grid representation performs well in classification and optical flow tasks. Specifically, the Event Voxel Grid is obtained by dividing event streams into a number of portions with equal temporal length and projecting all events within each portion onto individual image channels. If multiple events are projected onto the same pixel in a time channel, the intensity of that pixel accumulates based on the number of occurred events. Data of opposite polarities are handled separately and concatenated at the end of the processing step, doubling the total number of channels. Therefore, for an event camera with a spatial resolution of m by n, and an event stream of ΔT seconds, we can turn it into an Event Voxel Grid with $2k$ channels of size m by n, each corresponds to its polarity and time channel for $\Delta T/k$ seconds. EST [4] also proposed another event data representation named the Event Spike Tensor (EST), which is similar to the Event Voxel Grid structure. In the EST representation, the pixel intensities of each time channel are replaced by the normalized timestamps of the corresponding events. This representation produces the same data dimensions as the Event Voxel Grid method and preserves as much temporal information within each time bin as possible.

2.2 EventZoom

EventZoom [3] is a recently proposed neural network approach for event denoising and super-resolution. It's able to effectively remove noisy events and achieves SOTA super-resolution image reconstruction results while being 10x faster. The network takes in the noisy event data in low resolution and outputs its denoised version in high resolution, which both are represented by 3D Tensors. Before and after the network, there are processing steps to stack the raw events to tensors and re-distribute the output tensor to events.

As shown in Figure 1, EventZoom was built upon the 3D U-Net backbone and incorporated an E2I module to leverage the (2D RGB) image information. The U-Net first downsamples the event tensor with convolution layers to get a high-level feature

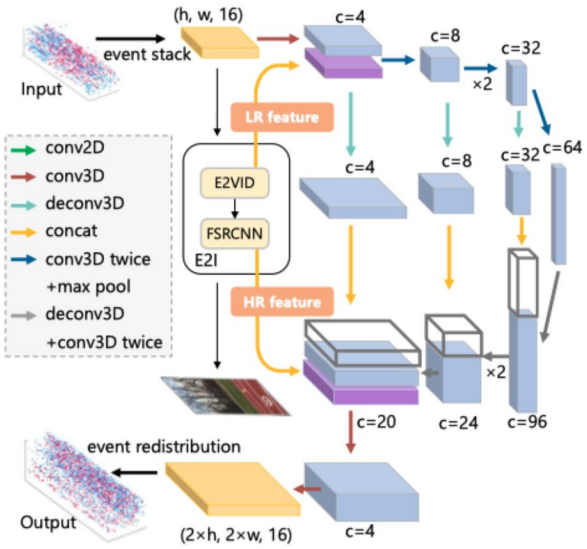

Figure 1: EventZoom-2× network architecture.

and then upsamples it with deconvolution layers until obtains the 2x feature map size. There are skip connections between the corresponding convolution and deconvolution layers with the same feature map size. The E2I module is a combination of an event-to-image reconstruction network E2VID[5] and an image super-resolution(SR) network FSRCNN[2] supervised by high-resolution ground-truth images.

3 Method

3.1 Data Processing

We plan to evaluate our approach on DVS gesture dataset [1]. This dataset consists of 11 hand gestures from 29 subjects under 3 illumination conditions. We add various types of noises to the spatial and temporal dimensions of the event data to simulate noisy input. As mentioned in Event Voxel Representation, an event stream sample can be converted to multi-channel image-like data. The noisy event samples are then transformed into event voxels for further processing. We also aim to explore different input data representations, such as the Voxel Grid and the Event Spike Tensor representations.

3.2 Baseline

As EventZoom [3] does not publicly release their code, we plan to implement our baseline model based on a 3D U-net backbone proposed by EventZoom [3]. The noisy event voxel data will be downsampled to high-level features and then deconvolved to high spatial resolution. Lastly, the high-resolution voxels will be assigned back to event streams, as also done in EventZoom.

3.3 Ours

We propose a new architecture that is modified from our baseline. We will introduce another convolution or MLP layer that encodes the temporal information within each voxel time channel. Then we add another decoder at the end of the network to assign voxel representation to the event stream, which is conditioned on the previously extracted temporal features. We believe that this method can increase the accuracy of denoising.

3.4 Evaluation Metrics

We evaluate the baseline and our approach's denoising performance using Signal to Noise Ratio (SNR) and Mean Squared Error(MSE) between the noisy event data and the recovered event data from the models.

4 Milestone

- Nov. 16 - Nov. 23: Set up the code base, data processing, evaluation pipeline, and implement the baseline
- Nov. 23 - Nov. 30: Implement and evaluate our approach
- Nov. 30 - Dec. 7: Prepare poster, report, and code submission

References

[1] Amir, A., Taba, B., Berg, D. J., Melano, T., McKinstry, J. L., di Nolfo, C., Nayak, T. K., Andreopoulos, A., Garreau, G., Mendoza, M., Kusnitz, J. A., DeBole, M. V., Esser, S. K., Delbrück, T., Flickner, M., and Modha, D. S. A low power, fully event-based gesture recognition system. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 7388–7397.

[2] Dong, C., Loy, C. C., and Tang, X. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision* (2016), Springer, pp. 391–407.

[3] Duan, P., Wang, Z. W., Zhou, X., Ma, Y., and Shi, B. Eventzoom: Learning to denoise and super resolve neuromorphic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 12824–12833.

[4] Gehrig, D., Loquercio, A., Derpanis, K. G., and Scaramuzza, D. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 5633–5643.

[5] Rebecq, H., Ranftl, R., Koltun, V., and Scaramuzza, D. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 3857–3866.