Flow-Event Autoencoder: Event Stream Object Recognition Dataset Generation with Arbitrary High Temporal Resolution

Minghai Chen[1,2]

[1]Guangdong Institute of Intelligence Science and Technology, Zhuhai, Guangdong, China
[2]Cognitive System Program, University of British Columbia, Vancouver, British Columbia, Canada
minghaipeterchan@outlook.com

Abstract

Event camera has unique advantages in high temporal resolution and dynamic range. However, due to the novelty of this hardware, there's a lack of large benchmark DVS event-stream datasets, including datasets for object recognition. In this work, we proposed an encoder-decoder method to augment event stream dataset from image and optical flow with arbitrary temporal resolution for object recognition task. We believe this proposed method can be generalized well in augmenting event stream vision data for object recognition and will help advance the development of event vision paradigm.

Introduction

Event-based cameras, such as Dynamic Vision Sensor (DVS), are innovative neuromorphic visual sensors that are becoming increasingly notable in the field of computer vision. Unlike conventional cameras that capture images at a fixed rate, event-based camera's pixels record changes of intensity asynchronously and it shows advantages in low energy consumption, high temporal resolution and high dynamic ranges (Gallego et. al. 2020). Event-based cameras have demonstrated promising results in several computer vision tasks such as optical flow estimation (Schnider et. al. 2023) and frame interpolation (Tulyakov et. al. 2021). Yet, the advancement of the event-based vision paradigm has been relatively slow, primarily due to the lack of large benchmark datasets, including object recognition datasets. These datasets are challenging to collect given the novelty of event-based cameras in daily life. Existing methods of generating or simulating event streams dataset suffer from: low efficiency and disturbance of refresh rate of screen, such as (Orchard et. al. 2015), poor temporal resolutions, such as (Zhu et. al. 2021, Hu, Liu and Delbruck 2021, and Gehrig et. al. 2020), lack of noise modelling (Gehrig et. al. 2020), require many manual-tune parameters (Hu, Liu and Delbruck, 2021) or are simply not designed for generating from images (Zhu et. al. 2021, Hu, Liu and Delbruck 2021). More effective methods of generating large datasets for object recognition from images are essential for the development of event-based vision paradigm. To fill these gaps, we propose an encoder-decoder (flow-event-flow) method for generating event-stream data from image and pre-defined saccades motion trajectory for large object recognition dataset with arbitrary high temporal resolution.

Background

Orchard et.al. (2015) started the generation of the event streams from static object images on display screen by applying saccades motion on DVS camera. It is currently the standard and most reliable method and it produced benchmark datasets such as N-MNIST, N-Caltech (Orchard et.al. 2015) and N-imageNet-1k (Kim et. al. 2018). Nonetheless, the display screen in this setup has a limited refresh rate, which could potentially affect its ability to generalize, as real-world objects do not have a refresh. Additionally, this method is not effective when applied to larger benchmark datasets with millions of images, as it requires mechanical set up of moving DVS to scan each image one at a time. The ESIM-Vid2e pipelines (Gehrig et. al. 2020, and Rebecq, Gehrig and Scaramuzza et. al. 2018) attempts to jointly render the DVS motion in a 3D scene and simulate process of DVS generates event spikes. However, it does not account for potential noise of DVS and relies on frame-upsampling algorithms to enhance temporal resolution, which may result in less precise temporal information. The v2e method (Hu, Liu and Delbruck, 2021) tries to model noises of DVS but shares the same issue of relying on frame-upsampling and has too many parameters on the simulated DVS camera for manual tune. Finally, EventGAN (Zhu et. al. 2021) attempts to directly generate event streams including accounts of noises but it is designed for video instead of images. Also, it can only generate fixed dimension data, thus it generates binned event-streams representation instead of raw event streams data. Our proposed method attempts to target the

gap of generating event streams from object image and resolve the issues from the existing methods to achieve: 1. Highly efficient and free of hardware setup during the actual generation process; 2. Good generalizability with no refresh rate issue and account of noise learning from actual DVS event streams data; 3. allowing arbitrary temporal resolution in generation, especially high temporal resolution.

Method

To generate event streams data from simulated saccades motion on object image, we propose to use a Vision transformer (ViT) encoder-decoder to learn to render events one-timestamp at a time from the optical flow extracted from the pre-defined motion trajectory and align the generated event streams to optical flow reconstruction task. We choose ViT over other architectures due to its self-attention mechanisms with advantages of global perceptual field information.

To start, we need a paired dataset of image-optical flow-event streams to train the model. We will construct this dataset by adopting the setup of a random saccadic moving DVS camera in (Orchard et. al. 2015) on the dataset of ImageNet (Kim et. al. 2021 and Deng et. al. 2009). Instead of displaying the object image on screen for DVS to record the event streams, we will print out the images from the non-reflective paper to prevent the issue of refresh rate of the display screen. Since the purpose is to train the model to capture the relationship between the motion trajectory and the event streams instead of constructing an entire dataset, we won't print the full dataset and will start with hundreds of images for resource conservation. The ground truth optical flow between any given timestamp will be extracted from the image's motion trajectory determined based on its relative position to the DVS camera during the saccadic motion of the DVS camera. With this data, we will build an encoder-decoder ViT and training pipeline as figure 1 shown. We will first pretrain a decoder network for optical flow estimation from the event streams, as have been successfully done in existing works (Schnider et. al. 2023). Next, at each time stamp *t*, we train the encoder to learn to generate event streams from the corresponding optical flow and Image. Two losses will be applied in this process: (1) L1 loss from ground truth event corresponding to the optical flow; (2) The predicted optical flow of pretrained decoder from the generated event streams; the ideas are to align the generated event streams to the optical flow estimation task: if the generated event streams are in high quality, the pre-trained decoder should be able to use it to reconstruct the ground truth optical flow.

The trained encoder can thus be used for generating event streams data of an object, given the image and the optical flow interpolate from its pre-generated motion trajectory of saccades motion. Users are free to generate event streams in any desired temporal resolution by tuning the interpolation time window when extracting optical flow.

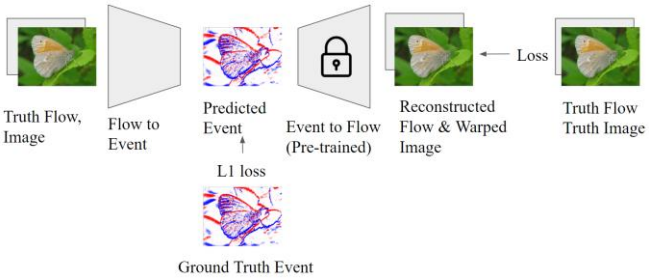

Figure 1. The proposed flow-event-flow reconstruction pipeline. The event to optical flow decoder is pre-trained and does not participate in parameter update. By doing this, we force the optical flow to event encoder to generate realistic event to align with the pretrained decoder to produce reliable optical flow and image reconstruction results.

Evaluation

To evaluate the validity of our methods, we will test whether our generated event streams can be aligned or out-perform other generation methods on the pretrain-transfer learning object recognition tasks. We will apply our proposed method and the ESIM-vid2e method (Gehrig et. al. 2020, and Rebecq, Gehrig and Scaramuzza et. al. 2018) to generate event streams data for ImageNet-1k (Deng et. al. 2009). Along with the event streams data that are mechanically generated in N-imageNet-1k (Kim et. al. 2021 and Deng et. al. 2009), we will pretrain ResNet34 models on each generated event streams data and compare the pretrained models' accuracies when transferring to another smaller event-based object recognition benchmark, such as the N-Caltech101 (Orchard et. al. 2015)

Discussion

We're expecting to see that the pretrained model on data from our proposed methods out-perform the methods of ESIM-vid2e (Gehrig et. al. 2020, and Rebecq, Gehrig and Scaramuzza et. al. 2018) and have comparable results to hardware saccades motion setup of (Orchard et. al. 2015) or even out-perform it in the object recognition task of N-Caltech101. Our proposed method provides a strategy to enrich object recognition dataset of event-based vision. It can potentially fuel the development of neuromorphic event-based vision paradigm by generating large, good quality datasets for large-scale pretraining and transfer learning, which has been prove to be vital for deep learning AI research (Chen et. al. 2023), and allowing these transferable models to be deployed on vital applications such as self-driving, high speed drone and robotics (Gallego et. al. 2020).

Conclusion

We proposed a ViT encoder-decoder pipeline for generating event-stream data from image and pre-defined saccades motion trajectory. Our method fills the gap of generating event streams data for object image and takes advantage of arbitrary temporal resolution and high efficiency of generation without hardware setup after training. We believe our proposed method can enrich object recognition dataset in event-based vision for large-scale pretraining and fuel the development of neuromorphic event-based vision paradigm.

References

Chen, F. L., Zhang, D. Z., Han, M. L., Chen, X. Y., Shi, J., Xu, S., & Xu, B. 2023. Vlp: A Survey on Vision-Language Pre-Training. *Machine Intelligence Research*, *20*(1), 38-56.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. 2009. Imagenet: A Large-Scale Hierarchical Image Database. In Proceedings of the Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers.

Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., ... & Scaramuzza, D. 2020. Event-Based Vision: A Survey. *IEEE transactions on pattern analysis and machine intelligence*, *44*(1), 154-180.

Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., & Scaramuzza, D. 2020. Video to events: Recycling Video Datasets for Event Cameras. In Proceedings of the Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers.

Hu, Y., Liu, S. C., & Delbruck, T. 2021. v2e: From Video Frames to Realistic DVS Events. In Proceedings of the Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers.

Kim, J., Bae, J., Park, G., Zhang, D., & Kim, Y. M. 2021. N-imagenet: Towards Robust, Fine-Grained Object Recognition with Event Cameras. In Proceedings of the international conference on computer vision. Institute of Electrical and Electronics Engineers.

Orchard, G., Jayawant, A., Cohen, G. K., & Thakor, N. 2015. Converting Static Image Datasets to Spiking Neuromorphic Datasets using Saccades. *Frontiers in neuroscience*, *9*, 437.

Rebecq, H., Gehrig, D., & Scaramuzza, D. 2018. ESIM: An Open Event Camera Simulator. In Proceedings of Conference on Robot Learning. Proceedings of Machine Learning Research.

Schnider, Y., Woźniak, S., Gehrig, M., Lecomte, J., Von Arnim, A., Benini, L., ... & Pantazi, A. 2023. Neuromorphic Optical Flow and Real-Time Implementation with Event Cameras. In Proceedings of the Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers.

Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., & Scaramuzza, D. 2021. Time Lens: Event-Based Video Frame Interpolation. In Proceedings of the Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers.

Zhu, A. Z., Wang, Z., Khant, K., & Daniilidis, K. 2021. EventGAN: Leveraging Large Scale Image Datasets for Event Cameras. In Proceedings of the International Conference on Computational Photography. Institute of Electrical and Electronics Engineers.