A Survey on Event Camera Simulators and Datasets for Optical Flow Estimation

Oisín Hageman[1]

Supervisor(s): Hesam Araghi, Nergis Tömen

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
July 3, 2024

Name of the student: Oisín Hageman
Final project course: CSE3000 Research Project
Thesis committee: Hesam Araghi, Nergis Tömen, Guohao Lan

Abstract

Computer vision tasks have shown to benefit greatly from both developments in deep learning networks, and the emergence of event cameras. Deep networks can require a large amount of training data, which is not readily available for event cameras, specifically for optical flow estimation. The need for simulating this data in a realistic, physics-driven manner is therefore crucial. This paper compares the state of the art event camera simulators on different criteria, including event timestamp modeling, performance under low illumination, bandwidth simulation, computation speed and various types of noise simulation. We also summarize the shortcomings of some commonly used optical flow event datasets. For generating high-quality, realistic events, The V2E and DVS-Voltmeter simulators have shown to produce the most accurate data.

1 Introduction

Event cameras, also referred to as dynamic vision sensors (DVS) are biologically-inspired sensors, that function differently from regular cameras. Key differences between bio-inspired event cameras and regular cameras include: higher temporal resolution, asynchronous pixels as opposed to synchronous frames, lower bandwidth, higher dynamic range (HDR) and only the ability to capture changes in brightness as opposed to absolute illumination on a pixel. Event-based vision can offer several benefits over vision with standard cameras [1].

The differences between event and regular cameras can provide opportunities for tasks like high-speed optical flow estimation, but algorithms for regular frame-based (RGB) cameras rely on different assumptions, e.g. synchronicity between frames, and can thus not be used. This discrepancy between assumptions also holds for available datasets. Standard, learning methods for optical flow estimation benefit largely from available synthetic datasets [2], but the same could not be said for the availability of event data with ground truths (until recently). The task of creating a dataset for training optical flow estimation models consists of several steps:

1. Use real video or render synthetic frames [2]
2. Simulate an event camera, to gather event data.
3. Collect the optical flow ground truth from the geometry + trajectories, or from the original video dataset if available.

This paper focuses on solutions for the second step, simulating the event cameras, as well as the availability of synthetic datasets and their shortcomings.

The main question we want to answer is: ”What features are most important in an event camera simulator, in the use case of training optical flow estimation models?” Our paper aims to provide the following two contributions: We compare the benefits and downsides of five available event camera simulators. And we summarize the available optical flow event datasets.

2 Related Work

Over the last years there have been numerous developments in event-based computer vision [1]. However, resources for high-level tasks such as recognition and classification are more numerous and broader in scope than of low-level tasks like optical flow estimation [3]. Key datasets for optical flow estimation include both MVSEC [4] and DSEC [5] which are event datasets with high quality optical flow ground truth. These datasets are specifically filmed on moving vehicles, and thus do not generalize well to other applications [2; 6]

This need for diverse event datasets, has caused the development of event camera simulators. Before these frameworks were created, a setup with a screen playing a video in front of an event camera was used in [7]. This method loses crucial characteristics of the event camera, namely the asynchrony of pixels, and HDR. The first event camera simulator was proposed by Kaiser et al. [8], but still stacks all events on single frame timestamps. ESIM [3] and its Vid2E extension [9], improve on this by using linear interpolation for timestamp sampling. V2E [10] is the best for training low brightness conditions, since it simulates shot noise prevalent at these conditions, that are not available in any other simulators. DVS-Voltmeter [11] is the newest simulator, and incorporates a model closer to the physics of the event pixels, and is thus able to more accurately simulate noise effects.

BlinkFlow [2] and [12] show that neural networks trained on synthetic event datasets benefit a lot from the generalisation of the data, where large-scale real data is missing.

3 Methodology

In the section 4, the benefits and downsides of five available event camera simulators are laid out. The key features we consider for the event camera simulators are:

- Timestamp Sampling
- Video Interpolation Method
- Accurate Low Illumination Scenes
- Noise Simulation
  - ’Circuit’ Noise, e.g. voltage/temperature noise and parasitic photocurrent
  - Temporal noise
- GPU Acceleration

An experimental setup for comparing the accuracy of an optical flow estimation model trained on synthetic event data versus real data can also be found in the discussion.

4 Event Camera Simulators

All event camera simulators discussed in this section, follow the following format: The input is either a frame-based video (Kaiser et al, Vid2E, V2E and DVS-Voltmeter), or a scene with trajectories to be rendered at a high framerate (ESIM).

Optionally, if the video is not of high enough framerate, the frames are interpolated. Next is the event simulation, where the simulator infers what events would have been generated between frames. These events are also given a timestamp based on a chosen sampling method. Finally there are multiple ways of simulating noise that an event camera would have generated in the data.

An overview of all features of the event camera simulators mentioned here can be found in Table 1. Since the method by Kaiser et al. is the first presented event camera simulator, it will be considered a baseline to compare the rest to.

4.1 ESIM and Vid2E

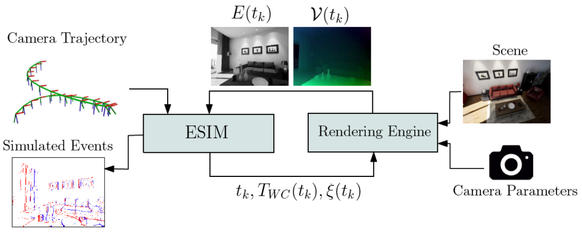


Figure 1: ESIM Architecture. ”At time t_k, ESIM samples a new camera pose $T_{WC}(t_k)$ and camera twist $\xi(t_k)$ from the user-defined trajectory and passes them to the rendering engine, which, in turn, renders a new irradiance map $E(t_k)$ and a motion field map $\mathcal{V}(t_k)$. The latter are used to compute the expected brightness change, which is used to choose the next rendering time t_{k+1}.” Figure from [3]

ESIM [3] is the first major improvement over the simulator from Kaiser et al. It does not directly take video data, but renders computer generated images instead. The simulator takes generally the same approach as [13], but improves upon the uniform sampling by using ”adaptive sampling” instead. Adaptive in this context means that the simulator will decrease the timestep, and thus increase the temporal resolution at moments when many events would be generated. This ”tight coupling” with the rendering engine makes the event data more realistic than its precursors. The only noise introduced by ESIM, is introduced by the variation of contrast threshold, which is modeled as Gaussian. Out of the simulators described here, ESIM is the fastest on CPU [2], and GPU acceleration was added after publication.

Vid2E [9] can be seen as an extension to ESIM. It uses the same event generation model as done by ESIM, but uses regular video as the input. To achieve the same high temporal resolution as achieved by the renderer, Vid2E instead uses SuperSloMo [14], a state-of-the-art video upscaling algorithm. SuperSloMo uses a convolutional neural network to predict optical flow in two directions and computes interpolated frames, to increase the temporal resolution of the video. After this, it follows the same method as ESIM.

4.2 V2E

V2E [10] has two major advantages over its predecessors. Firstly, the simulator adds a bandwidth restriction to its simulator, so events are not generated when a real event camera pixel would not due to the real bandwidth restriction. Secondly, V2E adds shot noise, which is especially prevalent in low illumination scenes. As such, V2E is the best event simulator for scenarios with low brightness.

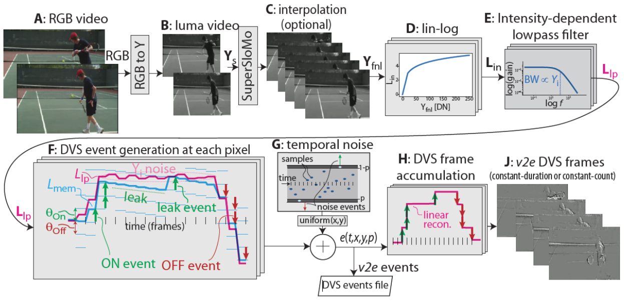
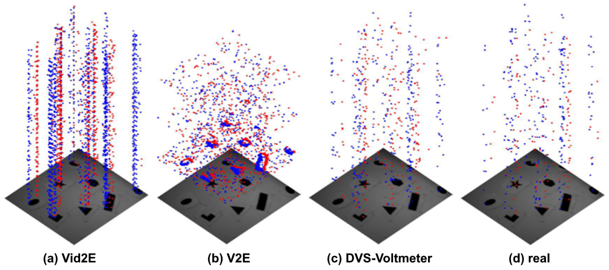


Figure 2: V2E Architecture. From [10]

One downside of V2E, compared to Vid2E and DVS-Voltmeter, is the timestamp sampling. As can be seen in Figure 3, the events are deployed at the discrete timestamps of the frames. As such the data loses the asynchronous pixel property of an event camera and makes it less realistic.

4.3 DVS-Voltmeter

DVS-Voltmeter [11] is the most recent event camera simulator. The algorithm is based on the specific characteristics of the circuitry behind event cameras. This approach aims to simulate noise effects in an event pixel as closely as possible, to make training data more suited to real world applications. The algorithm models voltage noise occurring in the circuit, random events caused by photon reception, and noise caused by temperature. Figure 3 qualitatively shows that their method makes the data resemble the real event data more closely, which will make overfitting less likely when used for training.



(a) Vid2E (b) V2E (c) DVS-Voltmeter (d) real

Figure 3: Comparison between event data from different simulators. Adapted from [11]

5 Datasets

5.1 Real Event Data

There are currently only 4 datasets available with real event camera data and optical flow ground truth. This can be attributed two factors: Event cameras are expensive, and research into event-based vision is not yet widespread, yet. As

Simulator	Type	Timestamp Sampling	Video Interpolation	Low Illumination	'Circuit' Noise	Temporal Noise	Bandwidth Limitation	GPU Acc.
Kaiser et al. [8]	Video Input	Discrete	-	n	n	n	n	n
ESIM [3]	Rendering	Linear	-	n	n	n	n	y
Vid2E [9]	Video Input	Linear	SuperSloMo	n	n	n	n	y
V2E [10]	Video Input	Discrete	SuperSloMo	y	n	y	y	y
DVS-VM [11]	Video Input	Brownian	SuperSloMo	n	y	n	n	n

Table 1: Comparison of Event Camera Simulator Features.

such, research facilities are not likely to invest in a event camera setup. Next to this, the setup for creating a real event dataset with optical flow ground truth, is not only difficult to produce, the resulting data is hard to combine, and not always of high quality. Subsequently making it less suitable for training neural networks. [2]

The available datasets are

- DVSFLOW16 [15]
- DVSMOTION20 [16]
- MVSEC [4]
- DSEC [5]

DVSFLOW16 and DVSMOTION20 use solely rotating camera's, and lack varying motion. They are also too small (5 and 4 scenes respectively) for training deep networks.

MVSEC and DSEC are specifically made for automotive purposes. For DSEC the setup is mounted on a car, and MVSEC on multiple vehicles. This makes these datasets not suited for use cases other than driving scenarios. The different sensors in the setup, e.g. event camera, video camera, etc. also are not perfectly calibrated with each other, adding another layer of inaccuracy to the data [3].

Another downside, is that of all four datasets, occlusion is not handled correctly [2].

5.2 Synthetic Event Data

BlinkFlow [2] is the only large-scale, diversiform optical flow event dataset to date. It was made using the BlinkSim suite, which is a rendering suite capable of creating event datasets with optical flow, and has integrated three of the camera simulators discussed: ESIM, V2E and DVS-Voltmeter. Their paper shows how their BlinkFlow training dataset performs better accuracy-wise when compared to MVSEC and DSEC training data on multiple models. Regrettably, the paper does not mention which of the three camera simulators was used for generating BlinkFlow. Nonetheless, their results show that synthetic data is a valid way of training deep networks for optical flow estimation.

6 Discussion

Originally, our research would include an experiment with a learning-based optical flow model. The aim was to compare a model trained on measured event data, with the same model trained on the simulated event data from the corresponding video, and see whether the simulated training data held up, or even exceeded the accuracy of the real data. The implementation of the experiment was abandoned due to time constraints, but the setup will still be provided here.

6.1 Experiment Setup

The experiment is in the same vein described included in section 6.2 of V2E [10], but for optical flow estimation, instead of classification. The setup makes use of the versatility of the MVSEC dataset, which contains real measured event data, regular video data, and optical flow ground truth [4]. The proposed steps for the experiment are as follows:

1. Use one of the event camera simulators, e.g. V2E, to convert the video frames of an MVSEC training partition into event data.
2. Train a state-of-the-art event-based optical flow network, such as E-RAFT [17] twice. Once on the real event data from MVSEC + optical flow, and once on the synthetic data + optical flow. The hyperparameters and seeds for the models need to be the same for a fair assessment.
3. Compare the accuracy between the models.

The experiment can be extended with multiple event camera simulators, to have an quantitative comparison for optical flow estimation between the different simulators as well. Like with the training of the model, we have to make sure the parameters between simulators, such as the upsampling rate, remains equivalent.

7 Responsible Research

In the interest of responsible research, this section is included to reflect on ethical aspects of data science.

While our paper lacks quantitative results from an experiment, the experiment setup laid out in the discussion aims to be reproducible. Our paper tries to adhere to the principles of FAIR data [18]. The paper is submitted to the Open Access TU Delft Repository (https://repository.tudelft.nl), where it can be viewed by anyone. Additionally, all simulators and datasets discussed this survey are open-source as well. These measures make it easier for researchers to make progress in the field of learning-based computer vision, which is especially important for niche or rising topics like event-based optical flow estimation.

8 Conclusion

This paper lays out different event datasets, and options for generating this data by the user. V2E and DVS-Voltmeter are among the best event camera simulators currently available.

However, as also mentioned in their respective papers [10; 11], adding the weaknesses, i.e. temporal shot noise for low illumination and accurate modeling of the event circuit respectively, could improve accuracy greatly. As future work, an event camera simulator that incorporates the features of both V2E and DVS-Voltmeter could make for even more realistic synthetic event data. Additionally, a quantitative study on the performance of differently generated event data, specifically for the task of optical flow estimation, is still missing in literature. The proposed experiment can be a step to finding out what characteristics of real event data are most important for training learning models.

References

[1] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

[2] Yijin Li, Zhaoyang Huang, Shuo Chen, Xiaoyu Shi, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Blinkflow: A dataset to push the limits of event-based optical flow estimation. 2023.

[3] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. *Conf. on Robotics Learning (CoRL)*, October 2018.

[4] Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, July 2018.

[5] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios, 2021.

[6] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. 2015.

[7] Elias Mueggler, Basil Huber, and Davide Scaramuzza. Event-based, 6-dof pose tracking for high-speed maneuvers. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2761–2768, 2014.

[8] Jacques Kaiser, J. Camilo v. Tieck, Christian Hubschneider, Peter Wolf, Michael Weber, Michael Hoff, Alexander Friedrich, Konrad Wojtasik, Arne Roennau, Ralf Kohlhaas, Rüdiger Dillmann, and J. Zöllner. Towards a framework for end-to-end control of a simulated vehicle with spiking neural networks. 12 2016.

[9] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2020.

[10] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events, 2021.

[11] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *ECCV*, 2022.

[12] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[13] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2), 2017.

[14] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation, 2018.

[15] Bodo Rueckauer and Tobi Delbruck. Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor. *Frontiers in Neuroscience*, 10, 2016.

[16] Mohammed Almatrafi, Raymond Baldwin, Kiyoharu Aizawa, and Keigo Hirakawa. Distance surface for event-based optical flow, 2020.

[17] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras, 2021.

[18] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 2016.