A Comparative Study of Model-based and Learning-based Optical Flow Estimation methods with Event Cameras

David Dinucu-Jianu (D.Dinucu-Jianu@student.tudelft.nl), Supervisor: Hesam Araghi, Responsible Professor: Nergis Tömen

Introduction

Optical flow estimation is a computer vision task that predicts motion in a video. **Event cameras**, with their **high temporal resolution**, are well suited for this task. For event cameras, motion must be predicted from the events they capture instead of frames.

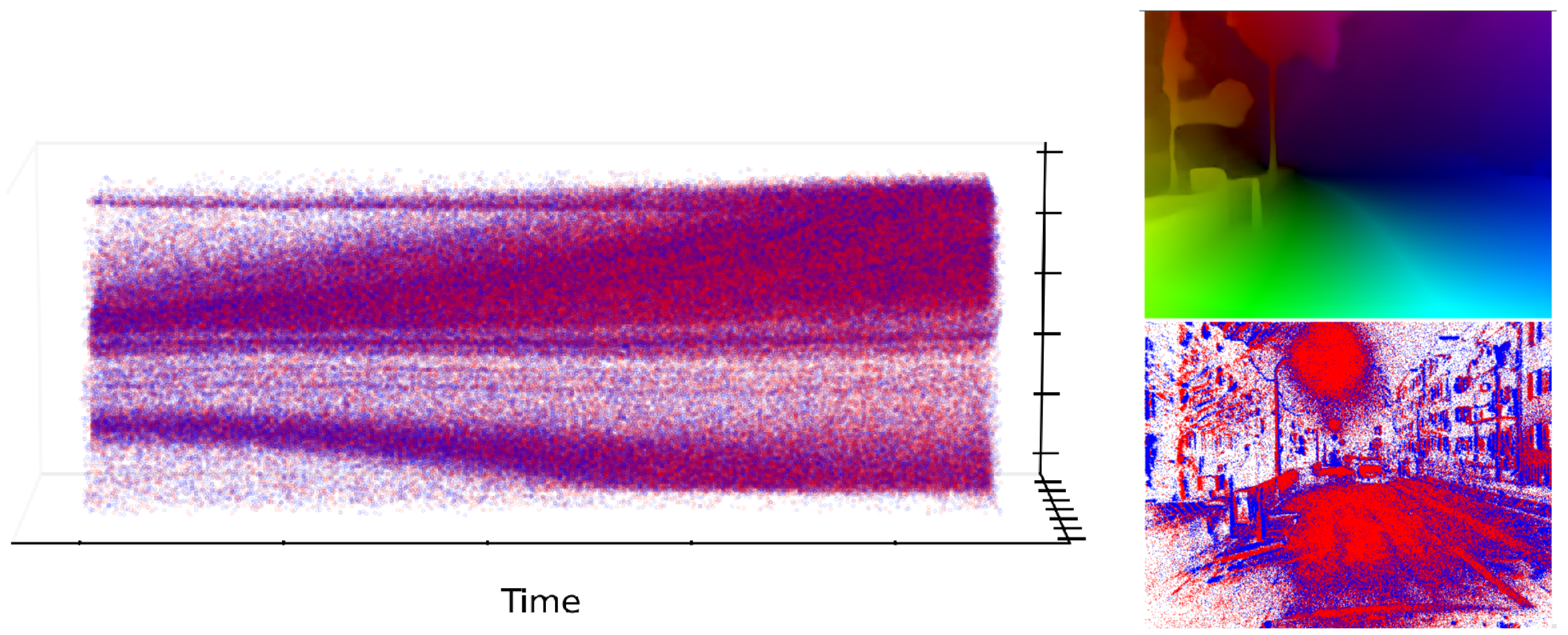

Figure 1. Stream of events, an event frame representation along with an optical flow visualization. Adapted from [3].

There are two major families of algorithms for event-based optical flow:

Model-based

- Generally optimization-based approaches
- Do not require training data
- Can be less computationally intensive

Learning-based

- Are neural network based
- Require large datasets of events
- Usually offer better accuracy

The new model-based approach **MultiCM** [6] achieved **state-of-the-art** accuracy on the **MVSEC** [8] dataset. However, it, along with another leading model-based method, Brebion et al. [1], significantly **underperformed** on the **DSEC** dataset [2].

Goal of the project

This study will compare the two approaches in terms of **accuracy** and **runtime** performance using the publicly available datasets **MVSEC** [8] and **DSEC** [2], aiming to inform the application of these algorithms. Additionally, we will investigate the performance gap observed on the DSEC dataset between learning and model-based approaches.

Benchmarks on the MVSEC and DSEC datasets

Table 1. Benchmark results on the MVSEC dataset. Learning-based methods are shown on top while model-based below. All results are reported from the respective papers. The performance of IDNet is shown for the 1/4 resolution version. It can be seen that model-based approaches perform better than learning-based ones.

	indoor_flying2		*indoor_flying3*		*outdoor_day1*	
	EPE ↓	$\%_{3PE}$ ↓	EPE ↓	$\%_{3PE}$ ↓	EPE ↓	$\%_{3PE}$ ↓
E-RAFT [3]	1.94	30.79	1.66	25.20	**0.24**	**0.00**
TMA [5]	1.81	27.29	1.58	23.26	0.25	0.07
IDNet [7]	-	-	-	-	0.31	0.1
Brebion et al. [1]	0.98	5.50	0.71	2.10	0.53	0.20
MultiCM [6]	**0.60**	**0.59**	**0.50**	**0.28**	0.30	0.10

Table 2. Benchmark results on the DSEC dataset. It can be seen that learning-based approaches perform significantly better as opposed to the results on MVSEC.

	EPE ↓	$\%_{1PE}$ ↓	$\%_{2PE}$ ↓	$\%_{3PE}$ ↓
E-RAFT [3]	0.788	12.742	4.74	2.684
TMA [5]	0.743	10.863	3.972	2.301
IDNet [7]	**0.719**	**10.069**	**3.497**	**2.036**
MultiCM [6]	3.472	76.57	48.48	30.855
Brebion et al. [1]	4.881	82.812	57.901	41.952

Runtime performance

Table 3. Runtime Comparison on DSEC Dataset. All benchmarks are performed on a laptop with an AMD Ryzen 7 5800HS CPU and an RTX 3060 Laptop GPU.

Model	CPU	GPU
E-RAFT [3] (1/8 Resolution, 12 iterations)	2.52s	130ms
TMA [5]	8.66s	246ms
IDNet [7] (4 iterations, 1/4 resolution)	7.70s	325ms
IDNet [7] (4 iterations, 1/8 Resolution)	2.23s	120ms
IDNet [7] (TID, 1 iteration, 1/8 Resolution)	530ms	24ms
MultiCM [6]	>10s	>10s
Brebion et al. [1]	63ms	39ms

Performance gap exploration

We will explore the gap in accuracy between model-based and learning-based approaches on DSEC. One theory, by Shiba et al. [6] is that learning-based approaches overfit to the predominant forward motion of DSEC thus artificially inflating their results.

Table 4. Comparison of retrained IDNet and TMA networks on the DSEC dataset along with MultiCM. It can be seen that despite not being trained on DSEC dataset the learning-based models outperform MultiCM.

	EPE ↓	$\%_{1PE}$ ↓	$\%_{2PE}$ ↓	$\%_{3PE}$ ↓
IDNet [7] (1/8 Resolution)	1.964	58.522	27.664	14.139
IDNet [7] (1/4 Resolution)	1.844	47.657	22.657	12.594
TMA [5]	1.938	51.618	21.111	9.693
MultiCM [6]	3.472	76.57	48.48	30.855

We retrained IDNet and TMA on the BlinkFlow [4] dataset, which includes a wider variety of motion types. We then evaluated this model on DSEC to check the claim of overfitting (Table 4).

Conclusion

We can draw the following conclusions about the two approaches:

1. **Model-based approaches**, provide the best runtime performance while maintaining good accuracy on datasets with small pixel displacements.
2. **Learning-based approaches** demonstrate superior accuracy on dynamic datasets but require GPUs to run in realtime.

Furthermore, the **accuracy gap** of model-based approaches on the DSEC dataset seems to stem **not only from the dataset's focus on forward motion** but also from **inherent limitations** of these algorithms.

References

[1] Vincent Brebion, Julien Moreau, and Franck Davoine. Real-time optical flow for vehicular perception with low- and high-resolution event cameras, 2021.

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

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

[4] 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.

[5] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. Tma: Temporal motion aggregation for event-based optical flow, 2023.

[6] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. *Secrets of Event-Based Optical Flow*. 2022.

[7] Yilun Wu, Federico Paredes-Vallés, and Guido C. H. E. de Croon. Lightweight event-based optical flow estimation via iterative deblurring, 2024.

[8] 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. 2018.