Unsupervised Event-based Optical Flow using Motion Compensation

Alex Zihao Zhu[0000−0002−2195−014X], Liangzhe Yuan[0000−0001−9206−1908], Kenneth Chaney[0000−0003−1768−6136], and Kostas Daniilidis[0000−0003−0498−0758]

University of Pennsylvania, Philadelphia PA 19104, USA

Abstract. In this work, we propose a novel framework for unsupervised learning for event cameras that learns to predict optical flow from only the event stream. In particular, we propose an input representation of the events in the form of a discretized 3D volume, which we pass through a neural network to predict the optical flow for each event. This optical flow is used to attempt to remove any motion blur in the event image. We then propose a loss function applied to the motion compensated event image that measures the motion blur in this image. We evaluate this network on the Multi Vehicle Stereo Event Camera dataset (MVSEC), along with qualitative results from a variety of different scenes.

Keywords: Event cameras, Unsupervised learning, Optical Flow

1 Introduction

Event cameras, such as in Lichtsteiner et al. [3], are a neuromorphically inspired, asynchronous sensing modality, which detect changes in log light intensity. The changes are encoded as events, $e = \{x, y, t, p\}$, consisting of the pixel position, x, y, timestamp, t, accurate to microseconds, and the polarity, p. The cameras provide numerous benefits, such as extremely low latency for tracking very fast motions, high dynamic range, and significantly lower power consumption.

Recently, several methods have shown that flow and other motion information can be estimated by 'deblurring' the event image [1, 5, 7]. For frame data, unsupervised optical flow methods such as [2, 4] have shown that neural networks can learn to predict optical flow from geometric constraints, without any ground truth labels.

In this work, we propose a novel input representation that captures the full spatiotemporal distribution of the events, and a novel unsupervised loss function that allows for efficient learning of motion information from only the event stream. Our input representation, a discretized event volume, discretizes the time domain, and then accumulates events in a linearly weighted fashion similar to interpolation. We train a neural network to predict a per-pixel optical flow from this input, which we use to attempt to deblur the events through motion compensation. During training, we then apply a loss that measures the motion blur in the motion compensated image, which the network is trained to minimize.

2 Method

We propose a novel input representation generated by discretizing the time domain. In order to improve the resolution along the temporal domain beyond the number of bins, we insert events into this volume using a linearly weighted accumulation similar to bilinear interpolation.

Given a set of N input events $\{(x_i, y_i, t_i, p_i)\}_{i=0,\ldots,N-1}$, we divide the range of the timestamps, $t_{N-1} - t_0$, which varies depending on the input events, into B bins. We then scale the timestamps to the range $[0, B-1]$, and generate the event volume as follows:

$$t_i^* = (B-1)(t_i - t_0)/(t_{N-1} - t_0) \tag{1}$$

$$V(x, y, t) = \sum_i p_i \max(0, 1 - |x - x_i|) \max(0, 1 - |y - y_i|) \max(0, 1 - |t - t_i^*|) \tag{2}$$

We treat the time domain as channels in a traditional 2D image, and perform 2D convolution across the x, y spatial dimensions.

Given optical flow for each pixel, $u(x, y), v(x, y)$, we propagate the events, with scaled timestamps, $\{(x_i, y_i, t_i^*, p_i)\}_{i=1,\ldots,N}$, to a single time t':

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + (t' - t_i^*) \begin{pmatrix} u(x_i, y_i) \\ v(x_i, y_i) \end{pmatrix} \tag{3}$$

We then separate these propagated events by polarity, and generate a pair of images, T_+, T_-, consisting of the average timestamp at that pixel, similar to Mitrokhin et al. [5]. However, by generating these images using interpolation on the pixel coordinates rather than rounding them, this operation is fully differentiable.

$$T_{\{+,-\}}(x, y, t') = \frac{\sum_i \max(0, 1 - |x - x_i'|) \max(0, 1 - |y - y_i'|) t_i}{N(x, y)} \tag{4}$$

where $N(x, y)$ is the number of events contributing to each pixel. The loss is, then, the sum of the two images squared, as in Mitrokhin et al. [5].

$$\mathcal{L}_{\text{time}}(t') = \sum_x \sum_y T_+(x, y)^2 + T_-(x, y)^2 \tag{5}$$

As we scale the flow by $(t' - t_i*)$ in (3), the gradient through events with timestamps closer to t' will be weighted lower. To resolve this unequal weighting, we compute the loss both backwards and forwards:

$$\mathcal{L}_{\text{time}} = \mathcal{L}_{\text{time}}(t_0) + \mathcal{L}_{\text{time}}(t_{N-1}) \tag{6}$$

We combine this loss with a spatial smoothness loss, $\mathcal{L}_{\text{smoothness}}$, applied to the output flow, with our final loss being a weighted sum of the timestamp loss and the smoothness loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{time}} + \lambda \mathcal{L}_{\text{smoothness}} \tag{7}$$

Our network consists of an encoder-decoder architecture, as defined in Zhu et al. [6].

	outdoor day1		indoor flying1		indoor flying2		indoor flying3	
dt=1 frame	AEE	%Outlier	AEE	%Outlier	AEE	%Outlier	AEE	%Outlier
Ours	**0.37**	**0.0**	0.59	**0.0**	1.02	3.2	0.89	2.5
EV-FlowNet	0.49	0.2	1.03	2.2	1.72	15.1	1.53	11.9
UnFlow	0.97	1.6	**0.50**	0.1	**0.70**	**1.0**	**0.55**	**0.0**
dt=4 frames	AEE	%Outlier	AEE	%Outlier	AEE	%Outlier	AEE	%Outlier
Ours	**1.23**	7.4	2.26	25.8	**3.92**	52.1	3.27	41.6
EV-FlowNet	**1.23**	**7.3**	**2.25**	**24.7**	4.05	**45.3**	3.45	39.7
UnFlow	2.95	40.0	3.81	56.1	6.22	79.5	**1.96**	**18.2**

Table 1: Quantitative evaluation of our optical flow network against EV-FlowNet and UnFlow. Average Endpoint Error (AEE) is computed in pixels, % Outlier is computed as the percent of points with AEE < 3 pix.

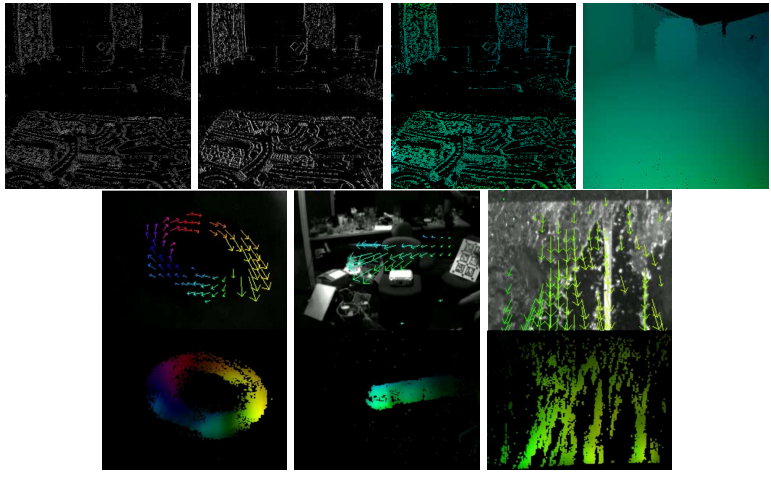

Fig. 1: Top: Result from MVSEC, left to right: blurred event image, deblurred image, predicted flow, ground truth flow. Bottom: Challenging scenes. Top images: sparse flow vectors on the grayscale image, bottom: dense flow output, colored by direction. Left to right: Fidget spinner spinning at 40 rad/s in the dark. Ball thrown quickly (the grayscale image does not pick up the ball). Water flowing outdoors.

3 Experiments

For all experiments, we train our network on the outdoor_day2 sequence from MVSEC [8], consisting of 11 mins of stereo event driving data. Each input to the network consists of 30000 events, with volumes with resolution 256x256 and $B = 9$ bins. The model is trained for 300,000 iterations, and takes around 15 hours to train on a NVIDIA Tesla V100.

For evaluation, we tested on the same sequences as in EV-FlowNet [6], and present a comparison against their results as well as UnFlow [4]. We convert

the output of our network, (u, v), into units of pixel displacement by the following scale factor: $(\hat{u}, \hat{v}) = (u, v) \times (B - 1) \times dt/(t_N - t_0)$, where dt is the test time window. From the quantitative results in Tab. 1, we can see that our method outperforms EV-FlowNet in almost all experiments, and nears the performance of UnFlow on the 1 frame sequences. As our event volume maintains the distribution of all of the events, we do not suffer from losing information as EV-FlowNet when there is a large motion. Our network also generalizes to a number of challenging scenes, as can be seen in Fig. 1.

4 Conclusions

In this work, we demonstrate a novel input representation for event cameras, which, when combined with our motion compensation based loss function, allows a deep neural network to learn to predict optical flow from the event stream only.

Acknowledgements

Thanks to Tobi Delbruck and the team at iniLabs for providing and supporting the DAVIS-346b cameras. We also gratefully appreciate support through the following grants: NSF-IIS-1703319, NSF-IIP-1439681 (I/UCRC), ARL RCTA W911NF-10-2-0016, and the DARPA FLA program.

References

1. Gallego, G., Rebecq, H., Scaramuzza, D.: A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In: IEEE Int. Conf. Comput. Vis. Pattern Recog.(CVPR). vol. 1 (2018)
2. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: European Conference on Computer Vision. pp. 3–10. Springer (2016)
3. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128x128 120 db 15μs latency asynchronous temporal contrast vision sensor. IEEE journal of solid-state circuits **43**(2), 566–576 (2008)
4. Meister, S., Hur, J., Roth, S.: UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In: AAAI. New Orleans, Louisiana (Feb 2018)
5. Mitrokhin, A., Fermuller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. arXiv preprint arXiv:1803.04523 (2018)
6. Zhu, A., Yuan, L., Chaney, K., Daniilidis, K.: EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In: Proceedings of Robotics: Science and Systems. Pittsburgh, Pennsylvania (June 2018). https://doi.org/10.15607/RSS.2018.XIV.062
7. Zhu, A.Z., Chen, Y., Daniilidis, K.: Realtime time synchronized event-based stereo. In: The European Conference on Computer Vision (ECCV) (September 2018)
8. Zhu, A.Z., Thakur, D., Ozaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multi vehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)