Unsupervised optical flow estimation of event cameras

Mark van den Berg
m.vandenberg-13@student.tudelft.nl
Bachelor Thesis CSE

The influence of training sets on model performance

Responsible professor: Nergis Tömen, supervisor: Hesam Araghi, examiner: Guohao Lan

Introduction

Event-based cameras are cameras that respond to changes in brightness, instead of capturing a set amount of frames every second.
They offer multiple advantages compared to regular cameras, namely **lower latency, high event rate, and high dynamic range**.
This different representation of vision compared to regular cameras means that regular computer vision algorithms that compare frames do not work on this data. This paper focuses specifically on optical flow estimation, the estimation of motion in a scene.

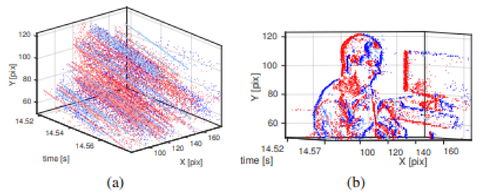

Fig 1. Contrast maximisation (a) shows events, both positive and negative in x,y,t space. Events are caused by movement in the scene (b) visualisation of the events from along the movement trajectory. Source: [1]

One of these optical flow algorithms shown in figure 1 is contrast maximisation consisting of the following steps:

- Create a linear flow prediction
- Lead all events back according to the predicted flow to a single time frame to create an image, also called the Image of Warped Events (**IWE**)
- Calculate the contrast of the IWE and choose flows to maximise the image contrast

The method **Taming Contrast Maximisation** uses an unsupervised neural network to predict flows. It uses the contrast of the IWE as a loss function to train the neural network.
It further improves upon the contrast maximisation framework by chaining **R** time frames together and calculating the flow iteratively over these time frames. These timeframes can even be split into halves **S** more times to create optical flow prediction over multiple timescales.
This leads to **nonlinear** flow predictions.

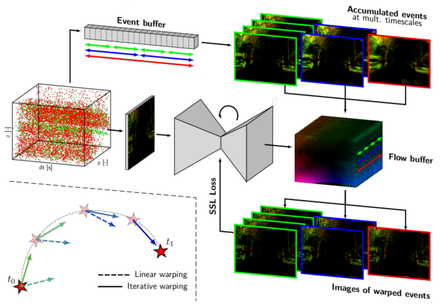

Fig 2. An overview of the Taming Contrast Maximisation algorithm. It consists accumulating events over multiple timeframes calculating their loss and updating the neural nerwork for these different losses. Source: [2]

This model in the original paper was trained on **DSEC** and **MVSEC**. MVSEC consists of a combination of driving and drone videos. DSEC consists of driving videos and is considered more difficult.

A new dataset was recently introduced called **BlinkFlow.** It consists of generated images of different objects and claims improved performance for supervised learning algorithms trained on it. But does it increase unsupervised learning performance?
Research question: "*What is the accuracy in terms of AEE, RSAT and FWL of the unsupervised model Taming Contrast Maximisation trained on BlinkFlow and evaluated on DSEC compared to the accuracy in terms of AEE, RSAT and FWL when trained on the DSEC dataset and evaluated on DSEC.*"

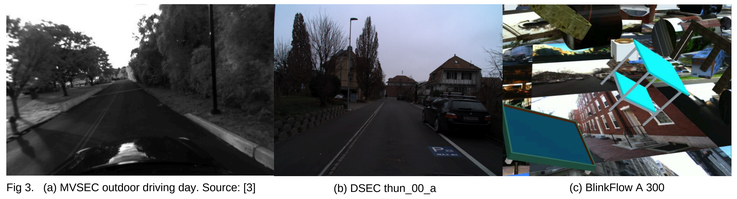

Fig 3. (a) MVSEC outdoor driving day. Source: [3] (b) DSEC thun_00_a (c) BlinkFlow A 300

Method

The hyperparameters R, S, and time frame, dt, are important hyperparameters that are dataset specific. Therefore we need to calibrate these parameters. Firstly multiple models are trained at different hyperparameters, then evaluated on the BlinkFlow dataset, and then these same models will be used to evaluate the DSEC dataset.
All models were trained between 6 to 14 hours and for a different amount of videos and epochs found in table 3. This was done due to time constraints of the research, and the high training times.

The models are evaluated with three metrics:

- **AEE**: average endpoint error, the average euclidean distance between the predicted flow and the ground truth flow.
- **FWL**: flow warp loss, a deblurring metric which aims to proxy accuracy when ground truth is not available.
- **RSAT**: another deblurring metric based on contrast to track accuracy when ground truth is not available.

	epochs trained	videos per epoch
dt = 0.01s, R = 2, S = 1	50	10
dt = 0.01s, R = 5, S = 1	1	20
dt = 0.01s, R = 10, S = 1	1	80
dt = 0.02s, R = 5, S = 1	2	10

Table 1 Amount training in epochs and video's per epoch per model.

Results

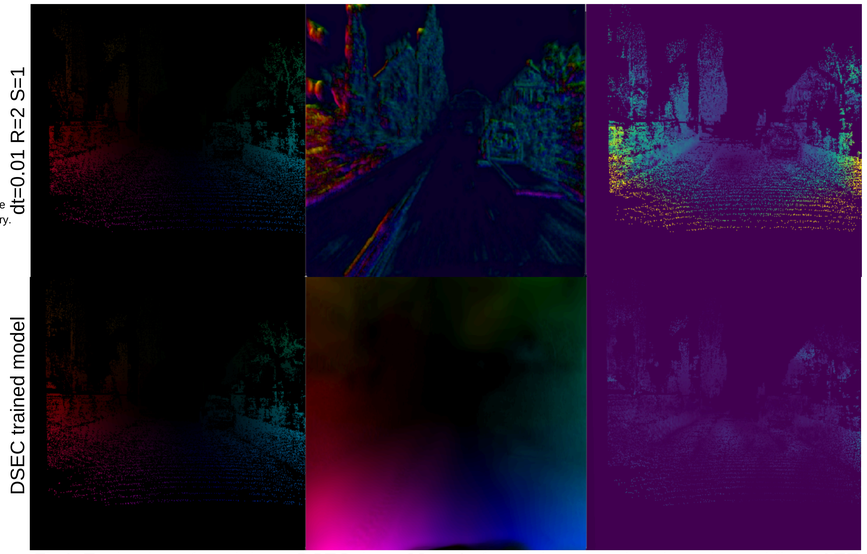

Fig4. Evaluation of models on DSEC thun_00_a. Left to right: ground truth, estimated flow, AEE

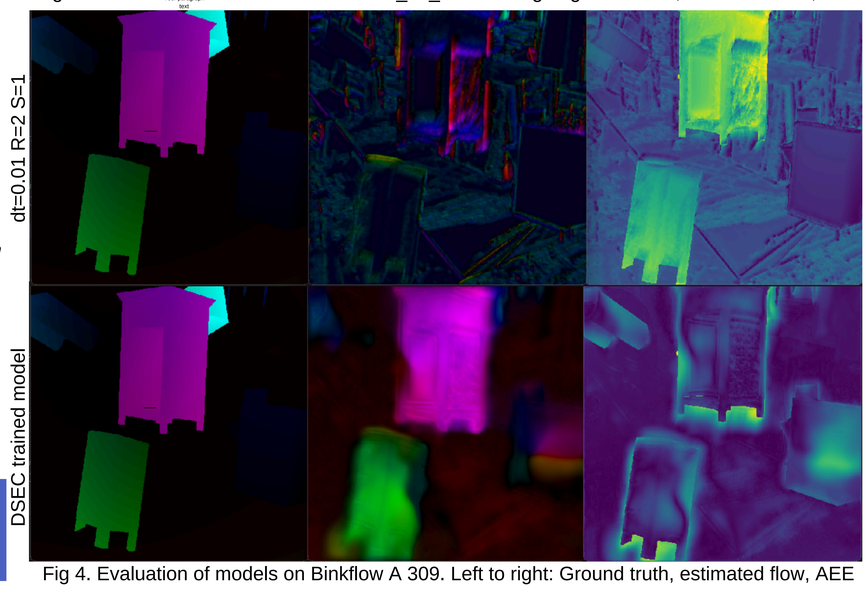

Fig 4. Evaluation of models on Binkflow A 309. Left to right: Ground truth, estimated flow, AEE

	AEE↓	FWL↑	RSAT↓	training time per video (s)	inference time (s)
dt = 0.01s, R = 2, S = 1	15.05	2.982	1.013	121	306
dt = 0.01s, R = 5, S = 1	16.06	0.997	0.999	485	286
dt = 0.01s, R = 10, S = 1	16.94	1.092	0.996	1504	290
dt = 0.02s, R = 5, S = 1	16.52	**4.402**	1.152	1293	296
DSEC best model	**6.17**	2.207	**0.753**	unknown	290
MVSEC best model	6.93	2.04	0.769	unknown	281

Table 2 comparing hyper parameters of models trained on BlinkFlow subsection A videos 0-10, evaluated on subsection A videos 300-309, best in bold, runner up is underlined. ↓ means lower is better, ↑ means higher is better. Runtime training calculated on HP Zbook with Quadro P2000. Runtime evaluation calculated on PC with GTX 1060TI

Discussion

The amount of training done on BlinkFlow by the Taming Contrast Maximisation models is too little to draw any definitive conclusions. This is also visible in figure 5, where the models with only one or two epochs of training were predicting flow in only one direction. This was thought to be because of overfitting, so the decision to increase the amount of training videos in the training set was made. This turned out to be counterproductive as it lead to fewer training epochs in the same training time.
Our best performing model was trained on only 10 videos of BlinkFlow which leads to 10s of video. This is only a very small subset of the BlinkFlow dataset of 3300 videos.
All videos in the training subset are part of BlinkFlow A, however, all parts of BlinkFlow seem to be generated the same way, so this should not influence the outcome.

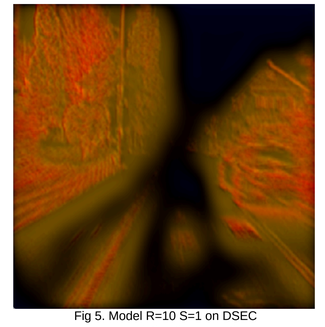

Fig 5. Model R=10 S=1 on DSEC always predicting the same flow

The models were not trained on the same amount of data which leads to unequal comparisons. This was because of limitations in processing power and time.

A better comparison could be comparing the models when trained on the same amount of videos.

Only very few metrics are used to evaluate the accuracy of the models as FWL does not seem to track AEE well.

Conclusion

The amount of training done by our models is too low to draw any conclusions about the generalisability of Taming Contrast Maximisation when trained on BlinkFlow.

The best performing pretrained model on DSEC, performs better than all our own trained models. This is most likely because of the larger amount of training done by the model pretrained on DSEC.

Our trained models seemed better at edge detection, while the pretrained model on DSEC seemed better at finding the flow of large surfaces. This could be due to the DSEC dataset containing only driving sequences which do not often encounter objects moving across the plane of vision in the same way that BlinkFlow does.

The DSEC pretrained model does not generalise well to the BlinkFlow dataset. The AEE drops around 350% when evaluating on BlinkFlow compared to DSEC. This could in part be due to the edge detection issues.

Another interesting find is that FWL does not seem to track accuracy well, but this is what the metric is designed to do. When evaluating on BlinkFlow the models with the highest FWL are our models, but these do not have a lower AEE. FWL is a metric which seems to not predict AEE well, although it does claim to be a proxy for accuracy.
The sample size of our results are too small to make any conclusions about the validity of FWL, but more research should be done concerning the validity of this metric.

Future work should also try to answer the research question posed in this research by training the Taming Contrast Maximisation model on a larger part of the BlinkFlow dataset for more epochs.

	AEE↓	FWL↑	RSAT↓
dt = 0.01s, R = 2, S = 1	9.31	1.962	1.102
dt = 0.01s, R = 5, S = 1	9.08	0.998	0.999
dt = 0.01s, R = 10, S = 1	9.94	1.022	0.989
dt = 0.02s, R = 5, S = 1	9.629	**2.547**	1.159
DSEC best model	1.758	1.188	0.870
MVSEC best model	**1.468**	1.280	**0.851**

Table 3 comparing models trained on BlinkFlow subsectionA video's 0-10, evaluated on subsection DSEC thun_00_a, best in bold, runner up is underlined. ↓ means lower is better, ↑ means higher is better

References:
[1] Guillermo Gallego, Henri Rebecq, and Davide Scara-muzza. A Unifying Contrast Maximization Framework for Event Cameras, with Applications to Motion, Depth, and Optical Flow Estimation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3867–3876, June 2018
[2] Federico Paredes-Vall´es, Kirk Y. W. Scheper, Christophe De Wagter, and Guido C. H. E. de Croon. Taming Contrast Maximization for Learning Sequential, Low-latency, Event-based Optical Flow, September 2023
[3] Hao Qu, Lilian Zhang, Xiaoping Hu, Xf He, Xianfei Pan, and Changhao Chen. Self-supervised Egomotion and Depth Learning via Bi-directional Coarse-to-Fine Scale Recovery