E-Stereo: Stereo Depth Estimation from Event Cameras

Yeongwoo Nam[1*], S. Mohammad Mostafavi I.[1*], Kuk-Jin Yoon[2] and Jonghyun Choi [1]
[1]Computer Vision Lab., GIST, South Korea
[2]Visual Intelligence Laboratory, Dept. Mechanical Engineering, KAIST, South Korea
yeong-oo@gist.ac.kr, mostafavi@gist.ac.kr, kjyoon@kaist.ac.kr, jhc@gist.ac.kr

Abstract

For stereo depth estimation using event cameras, we propose an end-to-end deep recurrent residual design taking advantage of a sequence of event representations that land before the location of ground-truth depth. From the sequential event inputs, we create a feature tensor used for aggregating and estimating regressed depth values. Our method outperforms the baseline method in all sequences and on average for all of the utilized metrics in the competition.

1. Introduction

The stereo event camera setting for depth estimation is not new and has been tackled in many ways including optimization-based methods and learning-based approaches [3]. MVSEC is one of the first datasets presenting events and images from a stereo set of event and image cameras on multiple indoor and outdoor scenes [4]. DSEC [1] is the most recent dataset that also has images in addition to events and is focused for driving scenes in a much larger scale. The events and images are from two different cameras which do have a small baseline. However as the calibration parameters and rectified images are provided we warp the images to the location of events once required.

2. Approach

The event stream cannot be delivered to convolutional neural networks and traditional computer vision algorithms, thus we stack them using the stacking based on time (SBT) method proposed in [2]. We use a sequence of three stacks reaching up to the intensity image timestamp that is also the ground-truth (GT) depth timestamp. For every GT depth, we extract the events up to 500 milliseconds preceding that location to keep our method causal. We divide the event stream to 3 equal sections and stack them using SBT which will create our sequence of events used as our input. Similar to the intensity images, the SBT event stacks are image-like representations. We use standard random image augmentations (rotate, flip, etc.) at training that remain the same per input sequence.

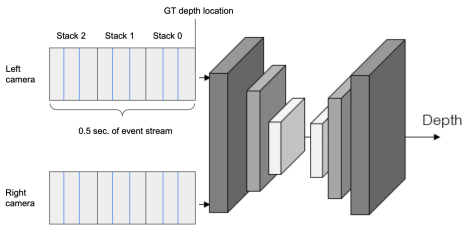

Figure 1. Abstract architecture. We use a sequence of input events and image frame from the left and right stereo pairs independently to create the fused input tensor to our depth estimation network.

An abstract representation of our network is presented in Fig. 1 We used a sequential design that leverages a residual recurrent sub-network for fusing the input sequence into a tensor. The tensor is used to extract depth features from each input pair. We aggregate the features to estimate the output regressed depth values. We train our method using the DSEC training data from scratch and submit the estimated depth results for evaluation on the test data in which the depth values are not provided.Our method is implemented with Pytorch, and we optimize our method using the end-point-error loss (L1 norm) on depth reconstruction.

References

[1] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. 1

[2] Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *International Journal of Computer Vision*, pages 1–21, 2021. 1

[3] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1527–1537, 2019. 1

* indicates equal contribution.

[4] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE RA-L*, 3(3):2032–2039, 2018. 1