LSMonRGBE: Learning-based Stereo Matching on RGB+Event Frames

Runqiu Bao
The University of Tokyo
bao@robot.t.u-tokyo.ac.jp

1 Introduction

This report is for attending the DSEC [1] competition of CVPR 2021 Workshop on Event-based Vision. The competition task is about estimating dense disparity from stereo event cameras and stereo global shutter cameras. Since both of the stereo pairs are respectively calibrated and rectified, the epipolar lines are horizontal. And therefore, estimating dense disparity is about finding pixel-level correspondences in the same row between the left and right camera in a stereo pair. Notably, final performance evaluation is on the left event camera, which means that the stereo global shutter cameras are only auxiliary.

The method recorded in this report can be illustrated as fig.1. Here, disparity estimation is treated as a traditional stereo matching problem, and the objective is to find optimal pixel-wise stereo correspondences by optimizing local patch differences of different disparity values for every pixels. Therefore, we aggregate the events within a certain time window into voxel representations. And to get more features for dense stereo matching, not only event representations but also the frames from the stereo global shutter cameras are remapped to the view point of the event cameras. Afterwards, the intensity images are stacked to the event representations along the channel axis, and stereo matching is performed on this comprehensive representation. Find details in Section 3.

Finally, it is noteworthy that, we are also aware that disparity estimation is more than just stereo matching. Because temporal continuity of data can be also utilized to smooth the current prediction results. However, temporal continuity is not considered in this method.

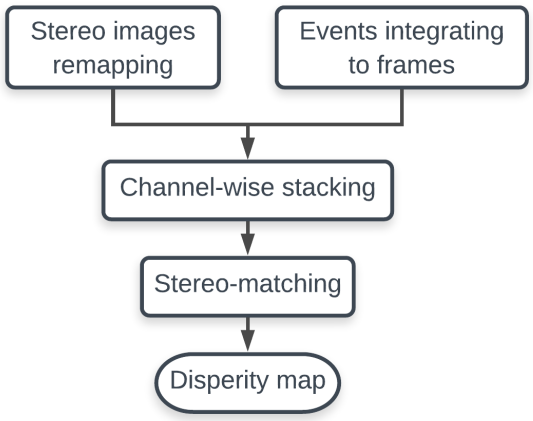

Figure 1: Method pipeline.

2 Results & Summary

The provided DSEC data set for training is sampled and divided into a training set of 1880 frames and a validation set of 716 frames. We use L1 loss for training and mean absolute error (MAE) as prediction performance metrics to evaluate model. The training includes 50 epoches, the metrics changed as fig.2 has shown. Finally, MAE has reached 0.4798 on the validation set.

In summary,

- Our approach is learning-based , specifically a learning-based stereo matching method is utilized.
- Events within a certain time window are integrated into voxel representations, whose time span is 30ms and contains 3 bins.
- Camera 0,1,2,3 are all used. Camera 0 and 3 are the stereo event cameras and camera 1, 2 are the stereo global shutter cameras.

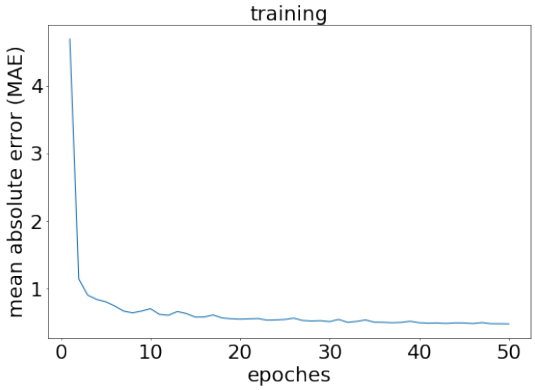

Figure 2: Method pipeline.

- For one prediction, the information used are the two stereo global shutter camera images, 30 ms of events symmetrical to the timestamp of the groundtruth disparity map.
- Part of the events are from the future. But in general, the method is causal, i.e. does not use information from future to predict at a given time.
- Learning-based stereo matching method is used, but the model is only trained on the training data of DSEC dataset.
- For data augmentation, RandomVerticalFlip is used. But it is not necessary. In fact, only 1888 frames sampled from the dataset are used for training and 716 frames for validation.

3 Details

This section includes some technical details about the method in fig.1.

3.1 Stereo Images Remapping

Using DSEC, first we want to align the views of global shutter cameras and event cameras pixel by pixel before combining them for stereo matching. Supposing cam0 and cam3 are the stereo event cameras and cam1, 2 the global shutter cameras. The remapping from cam1 to cam0 and cam2 to cam3 is as equation 1.

$$\begin{cases} T_{recti\_rectj} = \begin{bmatrix} R_{recti} & 0 \\ 0 & 1 \end{bmatrix} * T_{j\_i}^{-1} * \begin{bmatrix} R_{rectj} & 0 \\ 0 & 1 \end{bmatrix}^{-1} \\ X_{recti} = T_{recti\_rectj} * X_{rectj} \end{cases} \tag{1}$$

where $i = 0, 3$ and $j = 1, 2$, X represents a point in a specific camera coordinates, T is 4x4 homogeneous transformation matrix, R is 3x3 rotational transformation matrix and "rect" represents camera coordinates after rectification. Notably, this remapping is only possible since cam0,1 and cam2,3 are infinitively close to each other.

3.2 Events integrating to Frames

In DSEC, images from global shutter cameras have exposure time of around 15 ms. To align the features in intensity image with the features in events, time span of events is set to 30 ms and is symmetrical to the timestamp of the intensity image. The events are stacked into voxel representations as provided in the DSEC tool scripts. Each representation contains 3 bins, time span of each bin is therefore around 10 ms.

3.3 Channel-wise Stacking

The event representations of cam0,3 are stacked to the intensity images of cam1,2 in the channel direction. Therefore, the final input data for stereo matching are two 1 x 6 x H x W tensors, one from cam0,1 on the left and the other from cam2,3 on the right. And 6 channels include 3 channels from the RGB image and 3 bins from the voxel event representation.

3.4 Stereo Matching

We use a deep neural network model for stereo matching. It is modified from an existing NN model called GANet [2]. The basic principle is similar to traditional methods, which takes in two images and searches for spatially matching pixels to the pixels on the other.

Reference

[1] Gehrig Mathias, et al. "Dsec: A stereo event camera dataset for driving scenarios." IEEE Robotics and Automation Letters, 2021.

[2] Zhang Feihu, et al. "Ga-net: Guided aggregation net for end-to-end stereo matching." In proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.