16-833: Robot Localization and Mapping (Spring 2019)
Mid-Term Report

Fusing Intensity and Event-Based Frames for Deep Visual Odometry

Members: Ganesh Iyer (giyer), Abhay Gupta (abhayg), Suhit Kodgule (skodgule)

April 16, 2019

1 Introduction

Visual Odometry (the method of finding the continuous trajectory of the camera based on its egomotion) often suffers from faulty estimation in situations such as rolling shutter delays, high-speed motions, high dynamic range and sudden changes in illumination conditions. In this context, event-based cameras have shown great promise. However, a representation for event measurements is extremely crucial in deriving the egomotion of an event-based camera. Deep Learning has shown great promise in approximating ill-posed problems in computer vision with surprising efficiency. Therefore, we propose a novel deep learning framework that predicts the transformation [se(3)] output between pairs of frames. The key idea is the fact the prediction is supervised not only by the ground truth transformation between the frames but also by the per-pixel optical flow from the event stream.

2 Literature Survey

Traditional Visual Odometry methods follow a standard pipeline to estimate the pose transformation of the camera:

- From the image sequence, first a feature extraction method is used to find some invariant features in the scene.
- Once the features are detected, they are tracked into the next set of frames, while removing outliers.
- Based on the intrinsics of the camera and the correspondence between the set of frames the motion of the camera is estimated. In the case of direct methods, dense pixel errors are used (still relying on the camera intrinsics).
- Optionally, some scale estimation and local optimization are implemented to improve the trajectory output from the sequence.

In contrast to these type of methods, recently, various deep learning frameworks [1], [2], [3], [4] are being used to directly predict the transformation from the input frames. These methods have a significant advantage to traditional methods in the sense that they do not rely on feature engineering or any kind of

fine-tuning to recover pose. These frameworks are either directly supervised, or are weakly-supervised by other priors such as depth and optical flow.

A few methods in the current literature have tried to combine event-based camera streams with deep learning [5], [6], [7]. We extensively followed the literature and used a similar representation for our framework as well. Therefore, our method can be seen as an extension of [6] and [2], which can not only improve the reliance on the method in dynamic scenes but also make a deep learning system more reliable.

3 Discussion of Work Completed

Our current work has been focused on analyzing the requisite representation of the model. Since event streams are highly sparse at short timestamps, accurate representation is needed that can be averaged over the timestamps between the RGB frames. We use "Event-Based Time Surfaces" provided by [8] for our model. An Event can be composed as $e = x, t, p$, where x depends on the log intensity difference $log(I_{t+1}) - log(I_t)$, p is the polarity and t is the timestamp. The representation results in not only considering the event but also the age of the pixel based on the number of events at the point, encoding the motion at the pixel. We have also spent significant time on abstracting data, which we discuss in the next section.

3.1 Data Collection

In order to train our models and baselines, we are using the Multi-Vehicle Stereo Event Camera Dataset [9]. The dataset is quite rich, consisting of averaged events and grayscale images for a stereo camera pair. Stereo event data is collected from car, motorbike, hexacopter and handheld data, and fused with lidar, IMU, motion capture and GPS to provide ground truth pose and depth images. In addition, the dataset provides images from a standard stereo frame based camera pair (grayscale intensity images) for comparison with traditional techniques. An example is shown in Figure 1.

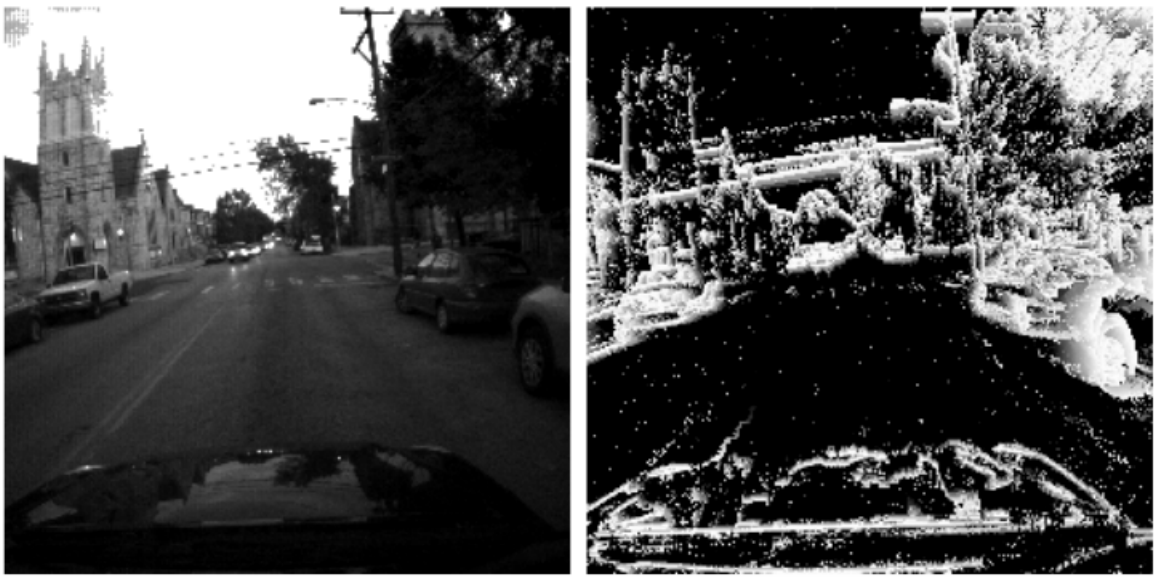

Figure 1: Example of a timestamp image. Left: Grayscale output. Right: Timestamp image, where each pixel represents the timestamp of the most recent event. Brighter is more recent

3.2 Baselines

We plan on comparing our method with both traditional VO approaches as well as deep learning based methods. In case of deep learning based approach, we will implement DeepVO [2] on MVSECD. DeepVO will solely utilize RGB frames, unlike our proposed method. It is an end-to-end learning method that utilizes recurrent convolutional neural networks (RCNNs) for predicting pose directly from a sequence of raw RGB images. The network learns efficient feature representations for VO using convolutional neural networks and models the sequential dynamics using Long Short-Term Memory (LSTM) networks. We will use pre-trained weights of the network previously trained on KITTI dataset for learning relevant feature representation. DeepVO will be an efficient baseline for evaluating the improvement observed by using event streams.

3.3 Proposed Method

The key idea is to use the weights learned by Ev-FlowNet [6] as the base network for extracting features from event streams. The image features that we propose will come from a modified implementation of the DeepVO network that uses a single intensity channel to learn the weights. We propose a fusion of the two features to output a single feature that represents the two streams. With this, we hope that the network can take advantage of the changes in inherent motion that event streams represent while learning the normal image features. These fused features at every timestep are passed to an LSTM which regresses the pose against the ground truth. The final trajectory which is learned compared against the ground truth to learn the model.

There are three key metrics which we will be evaluating our results against

- Absolute Translation Error (ATE)
- Relative Pose Error (RPE)
- L2 distance of $SE(3)$ Error

A figure of the proposed model is shown in Figure 2.

4 Technical Difficulties

One of the key challenges that we are facing is regarding the fusion of features. While various methods exist, we are still considering how to fuse interim features from a greyscale intensity model with the features from an event-based model.

Another technical difficulty that we have faced is due to data processing. Due to the use of rosbags, none of the event, intensity frames, and pose messages were synchronized. This was mitigated by using an approximate time synchronizer in ROS, so that all the messages were synchronized by the best possible approximate timestamp. Since the event messages still encode relative timestamps, the data is not affected by any synchronization changes.

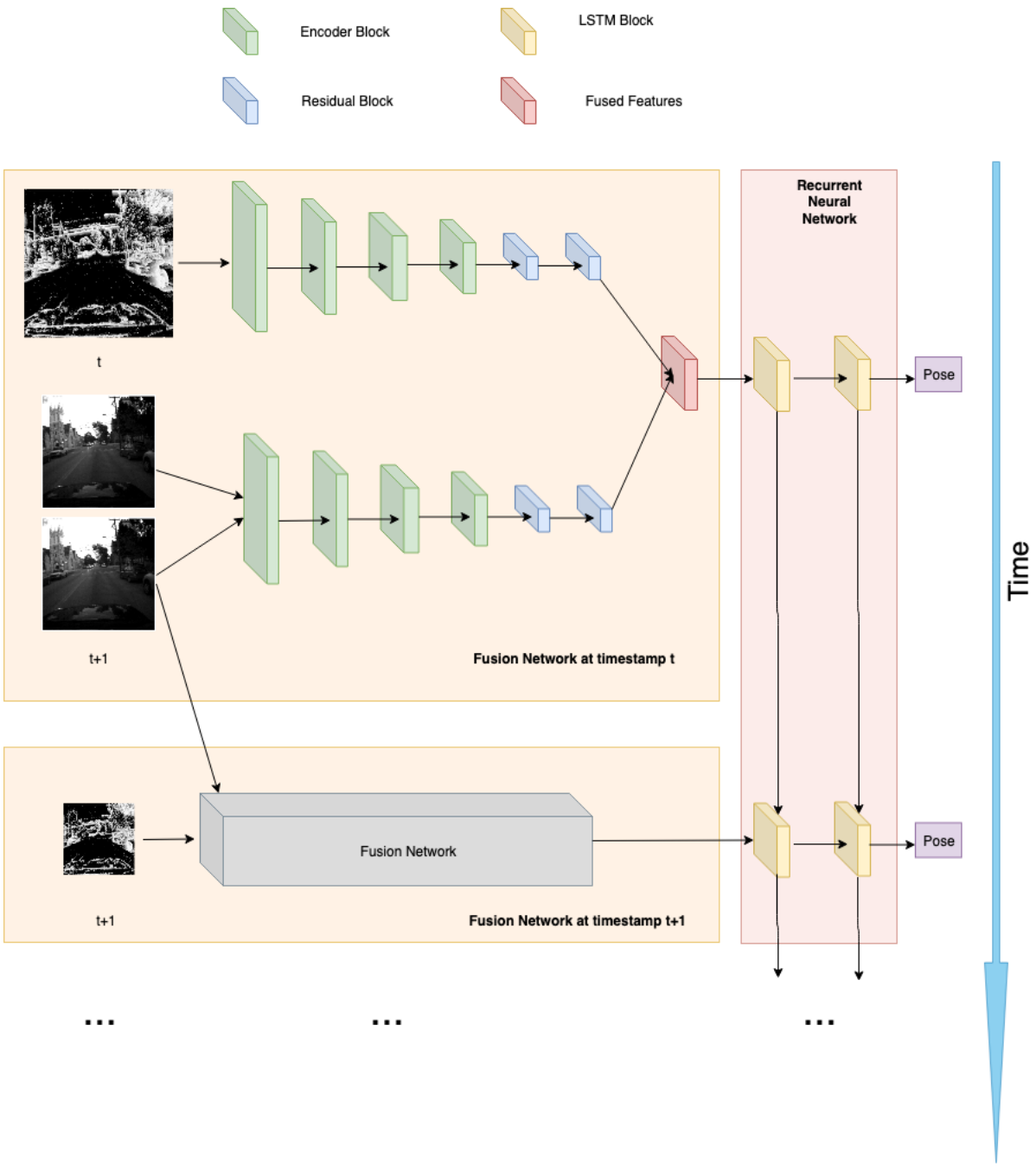

Figure 2: Proposed Fusion Model

5 Timeline

5.1 Current Situation with respect to Timeline

We are not making any changes to the original timeline. Data collection and identification of methods is completed. We are currently implementing the framework and the requisite baselines. We hope to achieve all goals without any major issues.

5.2 Changes in Timeline

No changes are proposed to the original timeline. To clarify, the subsequent timeline is as follows:

- 04/16 - Model reiteration - training/testing
- 04/22 - Benchmark Evaluations and Results
- 05/01 - Report and Presentation

References

[1] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[2] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050. IEEE, 2017.

[3] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7286–7291. IEEE, 2018.

[4] Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[5] Anh Nguyen, Thanh-Toan Do, Darwin G Caldwell, and Nikos G Tsagarakis. Real-time 6dof pose relocalization for event cameras with stacked spatial lstm networks. *arXiv preprint arXiv:1708.09011*, 2017.

[6] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.

[7] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018.

[8] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2017.

[9] 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 Robotics and Automation Letters*, 3(3):2032–2039, 2018.