Weekly Report (AUG 28, 2023 - Sep 3, 2023)

Zixun Xiong
Manning College of Information and Computer Sciences, UMASS Amherst
140 Governors Dr, Amherst, MA 01002
zixunxiong@umass.edu

Abstract

This week, I read a paper on dynamic object detection using CNN. At first, I followed the whole paper and related materials to get the basic idea. Then, I considered the limits of the current work while acknowledging the advantages of it. Despite the fact that the IMO detector part is quite well-defined, the dodging part especially for multi-object cases is quite naive, it only solves cases when the IMOs are targeting the drone. However, in real-world cases, when facing multiple IMOs this method didn't have good performance (76% success rate in the experiments). Then I read a paper under static scenes. This paper used graph spectral clustering to implement the IMO detection task that can determine k for the method automatically. I plan to test the performance of EVDogeNet on the bag file and real-world IMO detection test first, since the baseline model FAST didn't perform well even on this task when the hyperparameter was ill-conditioned. After that, I plan to add the Graph Clustering method to FAST for comparison to get a basic idea about what to improve and design the subsequent work accordingly.

1 EVDodgeNet: Deep Dynamic Obstacle Dodging with Event Cameras

This paper [1] introduces a deep learning-based solution to dynamic object avoidance. By using a series of shallow neural networks, researchers estimate both the ego-motion and the motion of independently moving objects (IMO). This method only needs the algorithm to be trained in simulation and can transfer to the real world without any fine-tuning or retraining. For the evaluation and testing, researchers test this approach in scenes with obstacles that have different shapes and sizes. The proposed method achieves an overall 70% success rate in scenes including objects that are not in the training set and scenes with low light.

1.1 Background and Significance

Independent Motion Detection and Ego-Motion Estimation. In visual Inertial Odometry (IMO), information from Inertial Measurement Units (IMU) is utilized to accomplish tasks like Simultaneous localization and mapping (SLAM). Works have been proposed by introducing event cameras to present a low-latency VIO algorithm to estimate ego-motion [2]. Most works before this paper focused on static scenes, which are rarely met in the real world. [3], however, mentioned that by carefully modeling, one can both estimate ego-motion and IMOs.

Image Stabilization. Recently, image stabilization is the most robust algorithm to make IMO more evident. Works inspired by this idea have been done on event cameras [4].

Obstacle avoidance on aerial robots. Works mentioned above are used to aid obstacle avoidance on aerial robots. An event camera has also been used for dodging high-speed obstacles in [5].

1.2 Challenges and Advantages

Although dynamic object detection using traditional cameras has been studied extensively, they are either of high latency or computationally expensive or don't fit for generalized missions including novel objects. Following [6], this paper applied deep learning to generalize the object detections for novel objects after being trained only on simulation.

EVDeBlurNet The input for event camera-aided VIO systems are event frames, which are generated by projecting event data to 2D arrays. This projection will cause misalignment [7]. Thereby, motion-compensations are needed. In this part, researchers use a shallow network named **EVDeBlurNet** to complete this task.

First, we get the event frames in equation 1 from a set of event data (shaped like $e = (\mathbf{x}, t, p)$), where **x** is the location of the triggered point in raw latent image I of the event camera.

$$\begin{aligned} E(x, \delta t)_{+} &= \sum_{t=t_0}^{t_0+\delta t} \mathbb{I}(x, t, p = +1) \\ E(x, \delta t)_{-} &= \sum_{t=t_0}^{t_0+\delta t} \mathbb{I}(x, t, p = -1) \\ E(x, \delta t)_{\tau} &= (\sum_{t=t_0}^{t_0+\delta t} \mathbb{I}(x, t, p = \pm 1))^{-1} \mathbb{E}(t - t_0) \end{aligned} \tag{1}$$

These generated event frames of given frequency δt are inputs of the model **EVDeBlurNet**, which is designed as an encode decoder structure. Thus the object model can remove stray events (which are generally noise) while preserving events corresponding to contours. Thereby, the loss function is defined as the equation 2.

$$\underset{E}{\text{argmin}} - C(\overline{E}) + \lambda D(E, \overline{E}) \tag{2}$$

$$C(E) \triangleq \mathbb{E}(||\mathbf{Var}(\nabla E)||) \tag{3}$$

$$D(E_1, E_2) \triangleq \mathbb{E}(||E_1 - E_2||_1) \tag{4}$$

Thus **EVDeBlurNet** can learn to denoise event frames, while images after the deblurring processing are not too far from the origin image by the second term in equation 2.

EVSegFlowNet. This network is aimed at detecting IMOs and dodging them accordingly. It will do both segmentation and optical flow optimization tasks. For the segmentation network, we use consecutive event frames to train a CNN model (outputting the probability for both background and foreground, which is the tag for the IMO), so that it can predict the foreground and the background segmentation. The loss function is defined as in equation 5. It's done on all pixels in even frames.

$$\underset{p_f}{\text{argmin}} \; - \mathbb{E}(\mathbb{I}_f \log(p_f) + \mathbb{I}_b \log(p_b)) \tag{5}$$

After being able to tell the difference between the background and IMOs, the next thing we need to deal with is the optical flow optimization. Considering that any unsupervised or self-supervised method can't predict the non-rigid optical flow (optical flows corresponding to the foreground regions) accurately since the unbalanced data where the number of foreground regions is far smaller than the background. However training on more data will cause the model to overfit, which is against the ideology of unsupervised training. To solve this problem, researchers combined optical flow optimization with the aforementioned segmentation labels (ground truth) by using a semi-supervised approach. This paper defines $\tilde{p_x}$ as the output of the second network, the formulation for this is defined in equation 6. Thus, the loss function is thereby defined in equation 7.

$$\tilde{\dot{p}}_x = \begin{cases} \dot{x}, & \text{if } \mathbb{I}_f(x) = 1 \\ 0, & \text{if } \mathbb{I}_b(x) = 1 \end{cases} \tag{6}$$

$$\begin{aligned} &\underset{\tilde{\dot{p}}_x}{\operatorname{argmin}}\, \mathbb{E}(D(W(E_t, \tilde{\dot{p}}_x) \circ \mathbb{I}_f, E_{t+1} \circ \mathbb{I}_f)) + \\ &\lambda_1 \mathbb{E}(||\tilde{\dot{p}}_x \circ \mathbb{I}_b||_1) + \lambda_2 \mathbb{E}(||\tilde{\dot{p}}_x \circ \mathbb{I}_b||_2^2) \end{aligned} \tag{7}$$

This network makes background flow zero fairly quickly compared to simple l_1 or quadratic penalty. The input to the EVSegFlowNet is $W(E_t, \tilde{H}_{4P_t})$ and E_{t+1}. To be mentioned, the definition of the function W can be found in [7].

1.3 Results and Experiments

synthetic datasets. Unlike existing datasets, researchers used a simulator to generate data with different textures and IMOs. Each scene consists of 3 different moving objects by varying random textures, and camera/object trajectories. They are all rendered 1000 frames per second at a resolution of 346×260 and a field of view $90°$

In the paper, researchers did 6 experiments to evaluate the performance of this algorithm. EVDogeNet did well (achieved 87% accuracy) when there exists prior information (e.g. size) of the target objects. However, when facing multi-objects (when the number of IMOs is more than 2), the performance decreased to 78%.

1.4 Disadvantages of the Paper

I wonder how this algorithm will react when IMOs are not targeting the drone. Intuitively, this will fail since the naive method mentioned in this paper is all based on the assumption that the IMOs are targeting the drone.

2 Moving Object Detection for Event-based Vision Using Graph Spectral Clustering

This paper introduced graph spectral clustering to IMOs detection using event cameras. What's more, it applied silhouette analysis to get the number of moving objects.

2.1 Background and Significance

Since even data generated by event cameras lack details like textures, it's hard to detect multiple objects in a single scene. Thus traditional methods used in RGB data can't be transferred to event data directly. Thus, it's a challenging task to derive IMOs information from the event camera. Moreover, considering the low cost of the event camera, it's preferable to explore robust algorithms for this data. Current methods [9][10] are either sensitive to noise or need to fine-tune a few parameters.

2.2 Challenges and Advantages

In general, this paper used kNN on the Laplacian Matrix to calculate k clusters, which stand for the moving object in this context. To help optimize this process, it used a metric named SC to indicate 'good segmentation'. This paper is quite simple can could be transferred to a framework like FAST [11].

3 Future Plan

I plan to test the performance of EVDogeNet on the bag file and real-world IMO detection test first, since the baseline model FAST didn't perform well even on this task when the hyperparameter was ill-conditioned. After that, I plan to add the Graph Clustering method to FAST for comparison to get a basic idea about what to improve and design the subsequent work accordingly.

References

[1] Sanket, Nitin J., et al. ”Evdodgenet: Deep dynamic obstacle dodging with event cameras.” 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.

[2] Vidal, Antoni Rosinol, et al. ”Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios.” IEEE Robotics and Automation Letters 3.2 (2018): 994-1001.

[3] Rebecq, Henri, Daniel Gehrig, and Davide Scaramuzza. ”ESIM: an open event camera simulator.” Conference on robot learning. PMLR, 2018.

[4] Stoffregen, Timo, et al. ”Event-based motion segmentation by motion compensation.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[5] Mueggler, Elias, et al. ”Towards evasive maneuvers with quadrotors using dynamic vision sensors.” 2015 European Conference on Mobile Robots (ECMR). IEEE, 2015.

[6] Falanga, Davide, Suseong Kim, and Davide Scaramuzza. ”How fast is too fast? The role of perception latency in high-speed sense and avoid.” IEEE Robotics and Automation Letters 4.2 (2019): 1884-1891.

[7] Gallego, Guillermo, Henri Rebecq, and Davide Scaramuzza. ”A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[8] DeTone, Daniel, Tomasz Malisiewicz, and Andrew Rabinovich. ”Deep image homography estimation.” arXiv preprint arXiv:1606.03798 (2016).

[9] Hinz, Gereon, et al. ”Online multi-object tracking-by-clustering for intelligent transportation system with neuromorphic vision sensor.” KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI, Dortmund, Germany, September 25–29, 2017, Proceedings 40. Springer International Publishing, 2017.

[10] Guang Chen et al. Neuromorphic vision-based multivehicle detection and tracking for the intelligent transportation system. Journal of Advanced Transportation, 2018.

[11] He, Botao, et al. ”Fast-dynamic-vision: Detection and tracking dynamic objects with event and depth sensing.” 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021.