EVENT CAMERA DEPTH ESTIMATION FROM EPIPOLAR PLANE IMAGES

Joshua D. Rego[1]*, Sanjeev Koppal*[2]*, Suren Jayasuriya*[1]

[1]Arizona State University, [2]University of Florida

ABSTRACT

Event-based cameras, known for their asynchronous detection of pixel brightness changes, have proven very successful in robotic vision and autonomous navigation applications for which depth estimation is crucial for enhancing downstream robotic localization and mapping. However, these tasks are often ill-posed and challenging, leading most state-of-the-art methods to rely on deep learning models for depth predictions. This reliance limits generalization to new scenes or camera parameters beyond the training datasets. In this paper, we present a conceptually different approach to depth estimation for event cameras inspired by single-shot lightfield capture and epipolar plane images (EPIs). We propose a robust sparse depth estimation pipeline based on Hough line detection on EPIs generated from event data. We demonstrate multi-view monocular event depth estimation by building a prototype with an event camera on a linear rail, demonstrating more accurate and generalizable performance compared to learning-based monocular and stereo event methods.

***Index Terms*—** Event-based Cameras, Depth Estimation, Computational Imaging

1. INTRODUCTION

Event cameras represent an innovative technology for visual perception with their low-power consumption, high dynamic range, and high temporal resolution. These sensors depart from conventional frame-based cameras by asynchronously detecting changes in brightness at the pixel level, termed "events". As a result, event cameras have found applications across diverse fields [1], including robotics, autonomous vehicles, surveillance systems, and augmented reality, where real-time sensing and robustness to motion are paramount.

Depth estimation stands as a crucial sub-problem in event-based vision systems, facilitating scene understanding and enabling applications such as obstacle avoidance, object tracking, and 3D reconstruction. Traditional depth estimation methods have primarily relied on stereo matching or monocular cues which can be challenging for event cameras. Recently, new deep learning techniques have shown promise in generating sparse or dense depth predictions, but often struggle to generalize and exhibit lower-accuracy on new scenes and camera parameters that are different than the datasets they are trained on. Further, they may require substantial computational resources for training and inference, limiting their applicability in environments that require reliable and accurate depth estimations.

In this paper, we propose a novel capture and depth estimation approach tailored specifically for event cameras, circumventing the limitations associated with learning-based methods. Our method leverages epipolar plane images (EPIs) derived from light field photography to extract depth information from event data. This approach offers simplicity and robustness, capable of extracting depth lines corresponding to sparse scene points directly from EPI images. We conduct evaluations of our proposed method against state-of-the-art event-based depth estimation, demonstrating its efficacy and robustness across various examples. Through the contributions in this paper, we aim to introduce EPIs for event sensing and pave the way for more efficient and reliable visual perception systems.

2. RELATED WORK

Monocular Depth Estimation from Events. Recent advancements in monocular depth estimation for event cameras have significantly enhanced the capabilities of event-based vision. Deep learning approaches have been shown to predict dense depth maps from event streams [2]. The spatio-temporal clustering method proposed by [3] uses real-time motion analysis on event-based 3D vision to extract depth information. Similarly, [4] presents a method for reconstructing 3D shapes using event data. Combining events with other modalities has also shown to be useful when combined with traditional frame-based methods in [5] to improve accuracy and robustness of monocular depth estimation through their developed recurrent asynchronous multimodal network, and with polarization information in [6] to estimate depth. Monocular methods have also been used for depth and 3D estimation with structured light [7, 8], and neural rendering fields (NeRFs) [9, 10].

Stereo Depth Estimation from Events. Stereo depth estimation for event cameras has similarly seen significant progress through various innovative approaches. Early work explored 3D reconstruction using stereo neuromorphic sen-

This work was supported by NSF CCF-1909663, NSF IIS-1909192 and a gift from Qualcomm.

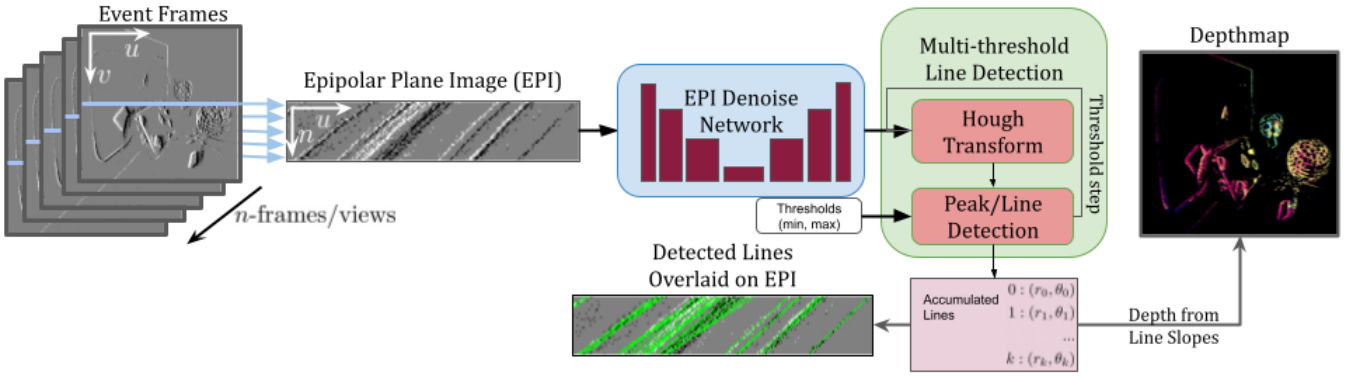

Fig. 1. Our proposed event depth estimation pipeline using multi-threshold Hough line detection to extract sparse depth maps. For n-frames, each row is stacked into an EPI, denoised, and passed through the multi-threshold line detection. The depth is then determined by the slope of the detected lines.

sors [11], Building on this, Andreopoulous et al. [12] developed a low-power, high-throughput, fully event-based stereo system, demonstrating the feasibility of efficient stereo depth estimation with event cameras. Zhu et al. [13] proposed a real-time, time-synchronized event-based stereo method. Ghosh and Gellego [14] introduced a fusion-based approach to enhance depth estimation accuracy using multiple event cameras. Deep learning-based methods for stereo depth estimation using event sequences have been proposed [15, 16]. Combining intensity images with event data has also been proposed for stereo methods [17, 18].

Some previous work has also been done with Hough Transform-based methods for event camera tracking. Glover et al. [19] use the Hough Transform to detect circular event objects for ball tracking, while [20] Tschopp et al. use a double Hough-based algorithm to detect vertical poles for tracking and mapping of railways.

3. PROPOSED METHOD

Epipolar Plane Image (EPI). A two-dimensional representation of a three dimensional scene can be obtained using the structure of an epipolar plane image (EPI) [21]. These are constructed by aligning the rows, or horizontal scan lines, of multiple images captured from different views along an epipolar line. For multiple viewpoints, linearly spaced along a plane, an EPI is formed by taking a row from each of the images and stacking them over each other. The image that this stack forms allows us to visualize disparities for a point in the scene across the different viewpoints in the form of diagonal lines corresponding to the pixel location of a scene point/feature through the shifted views. A point in the scene close to the camera will have a large disparity or line with a small slope, closer to a horizontal line, while a point far away from the camera will have a line with a large slope and be closer to a vertical line. The planar viewpoints that are stacked to form EPIs can be captured using multiple different methods. A 1-D camera array can simultaneously capture viewpoints of a scene, or similarly a single camera sliding along the planar rail can be used where viewpoints are temporal markers of the video as is the case for our experimental setup. Additionally, a theoretical single-sensor with a pinhole or microlens array can also be used.

Hough Line Detection. The Hough Line Transform [22] is a powerful technique in computer vision for detecting straight lines in images. The key insight behind the transform is to represent lines in a polar coordinate system, where each line is expressed as $r = x\cos(\theta) + y\sin(\theta)$. Here, r is the perpendicular distance from the origin to the line, and θ is the angle of this perpendicular. This representation avoids issues with vertical lines and allows for efficient line detection by converting edge points in the Cartesian plane into sinusoidal curves in the (r, θ) parameter space. The algorithm uses an accumulator array to count votes for each (r, θ) pair, and the cells with the highest votes indicate potential lines.

Overview of Depth Estimation Pipeline. Shown in Fig. 1, our pipeline initially loads the range of sequential frames to compute from. For each row index, y_i, the EPI is generated by stacking the row index from all sequential frames. From the 3-channel EPI we process the positive event (red) and negative event (blue) channels as separate grayscale EPIs to create larger separation of positive and negative event lines that may be too close together. Both EPIs go through denoising and morphological functions of dilation and erosion to increase the prominence of the lines. Each EPI is then sent to the multi-threshold Hough line detection function and depth is extracted from the detected lines and assigned to a depthmap.

Denoising Network. Event noise in captured scenes can cause errors for line detection in the EPIs. This is caused in two ways: additional events that are not from objects in our scene, and missing events from our object edges causing

Fig. 2. Experimental setup with Prophesee Mk3 event camera on a motorized linear rail. The rail is programmed to translate the camera a short distance to capture different viewpoints along a plane to form the EPIs.

inconsistencies in the lines themselves. Reducing noise of the first type is much more beneficial as the minimum threshold limit can then be reduced to detect the more inconsistent sparse lines.

To denoise, we use a simple UNet-based network that removes noise from the EPI while preserving the lines. The network is trained on simulated EPIs with inconsistent random lines as ground truth and added salt-and-pepper noise for the input. The output generates a mask to filter through only the lines, and the network is trained using $L1$-distance between the filtered EPI and the ground truth.

Multi-Threshold Line Detection. We initialize the algorithm at the upper bound of a threshold range, passing the initial EPI through Hough line detection, which returns an array of detected lines (r, θ). These lines are added to a cumulative list and remove from the EPI. The updated EPI is processed again with the next lower threshold, adding new detected lines to the list and subtracting them from the EPI. This process continues until the lower bound of the threshold range is reached.

Higher thresholds more accurately detect closer object depths, so detecting and removing these lines early helps prevent errors at lower thresholds, ensuring each feature corresponds to a single line and depth. The upper bound of the threshold range is typically defined to be a little higher than the number of views/frames used for the EPI, while the lower bound is roughly half of the upper bound.

Depthmap Generation. From the lines detected by multi-threshold Hough detection, depth is determined by calculating each line's slope and assigning it to the depth map at the pixel corresponding to the line's x-intercept in the first frame and row r.

4. DATA AND IMPLEMENTATION

We capture real event scenes using the Prophesee Mk3 camera (1280×720 resolution, 5mm C-mount lens) for depth estimation analysis. The camera is mounted on a motorized linear rail, programmed via Arduino for translation between two positions, and synchronized with event camera software for capture. This setup is shown in Fig. 2. Event data is captured using Metavision Studio without noise filtering and exported as .txt files, where each event is represented as (t, x, y, p) where t is time, (x, y) is spatial location, and $p = \pm 1$ is the polarity of the event. The data is later formatted as accumulated 2D frames for our method or voxel grids for comparison methods.

After forming and denoising the EPI, we apply dilation and erosion to enhance tracklines before line detection. Once depth values are extracted from the EPI lines and placed on the depth map, we perform post-processing with dilation, a median blur (kernel size 3), and filter depth values based on event locations. This ensures accurate depth estimation at the correct locations and consistency when compared to dense depth methods.

Baselines. We compare against [2] for a monocular event-based camera method, and [16] for a stereo method. Both methods are deep learning-based and produce dense depthmaps. For consistency, dense depth is filtered using events as described in the previous section.

5. EXPERIMENTAL RESULTS

We present depth estimations for an example real scene captured with the Prophesee Mk3 in Fig. 3, featuring a teacup in the foreground, a drink can in the background, and a Stanford bunny in the middle. To assess performance, we addition-

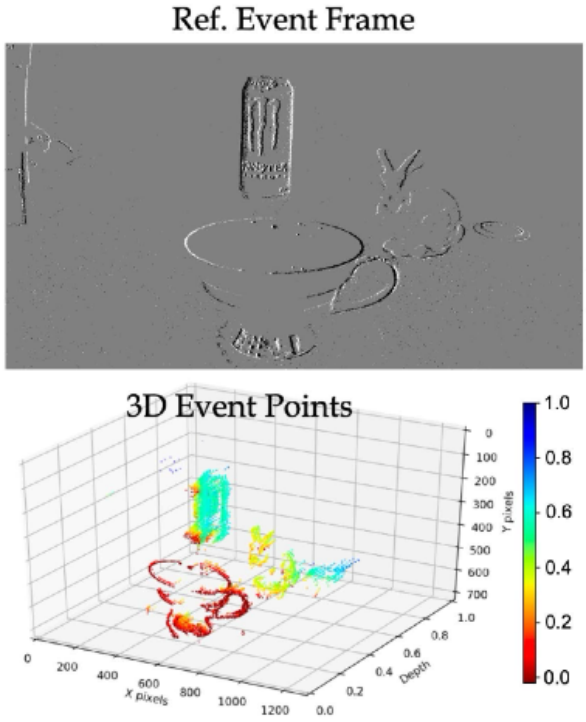

Fig. 3. EPI depth on a real scene: reference frame (top) and resulting 3D event point cloud (bottom). Relative depth is reported in normalized units from 0 to 1.

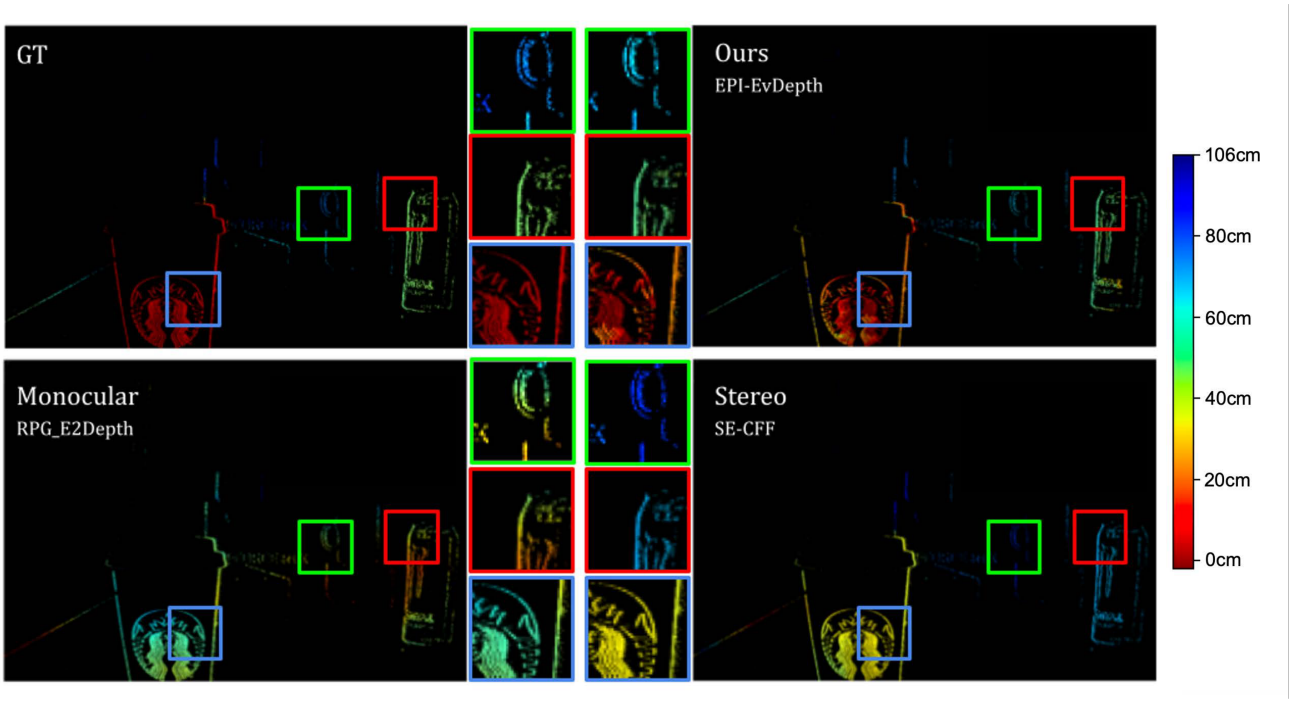

Fig. 4. Qualitative comparisons on real data for our method vs. a monocular (RPG-E2Depth) and a stereo (SE-CFF) depth estimation methods. Quantitative results for this scene are reported in Table 1.

ally compare our method with deep learning-based monocular and stereo depth estimation methods through qualitative and quantitative analysis.

In Fig. 4, depth estimates for all three methods are shown. The monocular method incorrectly estimates object depths, while the stereo method provides better object separation but incorrect depths. Our method delivers more accurate depth estimates for the objects, though with some errors compared to the ground truth. We also measure quantitative results against the baseline methods averaged across several real scenes. Conversion from relative normalized depth to true metric depth was done using a 3D scan of the scene for all methods. These are shown in Table 1 with MAE, MSE, and RMSE error metrics to show that our method performs noticeably better than both baseline methods.

Table 1. Quantitative comparisons on real data for our method vs. a monocular (RPG-E2Depth [2]) and a stereo (SE-CFF [16]) depth estimation methods

	Monocular	Stereo	Multi-view Monocular
	RPG-E2Depth	SE-CFF	EPI-EvDepth (Ours)
MAE	40.63cm	29.94cm	11.50cm
MSE	18.47cm	9.41cm	2.66cm
RMSE	44.25cm	31.59cm	16.79cm

6. DISCUSSION

In this paper, we demonstrated that depth estimation for event cameras using epipolar plane images (EPIs) with our multi-threshold Hough Transform-based pipeline. However, there still remains limitations for our current approach. EPI-based depth estimation remains sparse, preventing dense depth estimation. Our EPI acquisition is horizontal or vertical, and would not work for general camera trajectories. There is a tradeoff in the speed of the rail, event acquisition, and the speed of objects that can be reconstructed. While event cameras can effectively achieve KHz frame rates, the speed of the rail can only move a few centimeters a second. That is why we require static scenes in this experimental prototype.

However, if a theoretical light-field based event camera with microlenses above the sensors could be fabricated, then EPI images could be extracted for dynamic scenes. All these challenges remain future work, as well as expanding the method to incorporate full event-based light fields using 2D motion or optical arrays for applications in refocusing, novel view synthesis, and/or single-shot monocular depth estimation.

7. REFERENCES

[1] Guillermo Gallego et al., "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.

[2] Javier Hidalgo-Carrio, Daniel Gehrig, and Davide Scaramuzza, "Learning monocular dense depth from events," *International Conference on 3D Vision*, 2020.

[3] Stephan Schraml and Ahmed Nabil Belbachir, "A spatio-temporal clustering method using real-time motion analysis on event-based 3d vision," in *IEEE Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 57–63.

[4] Alexis Baudron, Zihao W Wang, Oliver Cossairt, and Aggelos K Katsaggelos, "E3d: event-based 3d shape reconstruction," *arXiv preprint arXiv:2012.05214*, 2020.

[5] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrio, and Davide Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *IEEE Robotic and Automation Letters. (RA-L)*, 2021.

[6] Manasi Muglikar, Leonard Bauersfeld, Diederik Moeys, and Davide Scaramuzza, "Event-based shape from polarization," in *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, Jun 2023.

[7] Manasi Muglikar, Diederik Paul Moeys, and Davide Scaramuzza, "Event guided depth sensing," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 385–393.

[8] Huijiao Wang, Tangbo Liu, Chu He, Cheng Li, Jianzhuang Liu, and Lei Yu, "Enhancing event-based structured light imaging with a single frame," in *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2022, pp. 1–7.

[9] Simon Klenk, Lukas Koestler, Davide Scaramuzza, and Daniel Cremers, "E-nerf: Neural radiance fields from a moving event camera," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1587–1594, 2023.

[10] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik, "Eventnerf: Neural radiance fields from a single colour event camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4992–5002.

[11] João Carneiro, Sio-Hoi Ieng, Christoph Posch, and Ryad Benosman, "Event-based 3d reconstruction from neuromorphic retinas," *Neural Networks*, vol. 45, pp. 27–38, 2013.

[12] Alexander Andreopoulos, Hirak J Kashyap, Tapan K Nayak, Arnon Amir, and Myron D Flickner, "A low power, high throughput, fully event-based stereo system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7532–7542.

[13] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis, "Realtime time synchronized event-based stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 433–447.

[14] Suman Ghosh and Guillermo Gallego, "Multi-event-camera depth estimation and outlier rejection by refocused events fusion," *Advanced Intelligent Systems*, 2022.

[15] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, and M. Hirsch, "Learning an event sequence embedding for event-based deep stereo," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[16] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi, "Stereo depth from events cameras: Concentrate and focus on the future," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Patter Recognition*, 2022.

[17] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi, "Event-intensity stereo: Estimating depth by the best of both worlds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4258–4267.

[18] Jinjin Gu, Jinan Zhou, Ringo Sai Wo Chu, Yan Chen, Jiawei Zhang, Xuanye Cheng, Song Zhang, and Jimmy S Ren, "Self-supervised intensity-event stereo matching," *arXiv preprint arXiv:2211.00509*, 2022.

[19] Arren Glover and Chiara Bartolozzi, "Event-driven ball detection and gaze fixation in clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2203–2208.

[20] Florian Tschopp, Cornelius Von Einem, Andrei Cramariuc, David Hug, Andrew William Palmer, Roland Siegwart, Margarita Chli, and Juan Nieto, "Hough ̂2 map–iterative event-based hough transform for high-speed railway mapping," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2745–2752, 2021.

[21] Robert C Bolles, H Harlyn Baker, and David H Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[22] Paul VC Hough, "Method and means for recognizing complex patterns," Dec. 18 1962, US Patent 3,069,654.