6-2024

Poster: Towards efficient spatio-temporal video grounding in pervasive mobile devices

Dulanga Kaveesha WEERAKOON MUDIYANSELAGE
Singapore Management University, mweerakoon.2019@phdcs.smu.edu.sg

Vigneshwaran SUBBARAJU
Institute for High Performance Computing

Joo Hwee LIM
Institute for Infocomm Research (I2R), joohwee@i2r.a-star.edu.sg

Archan Misra
Singapore Management University, archanm@smu.edu.sg

Citation

WEERAKOON MUDIYANSELAGE, Dulanga Kaveesha; SUBBARAJU, Vigneshwaran; LIM, Joo Hwee; and Misra, Archan. Poster: Towards efficient spatio-temporal video grounding in pervasive mobile devices. (2024). *MOBISYS '24: Proceedings of the 22nd Annual International Conference on Mobile Systems, Minato-ku, Tokyo Japan, 2024 June 3-7*. 648-649.

Available at: https://ink.library.smu.edu.sg/sis_research/9219

Poster: Profiling Event Vision Processing on Edge Devices

Ila Gokarn
Singapore Management University
Singapore
ingokarn.2019@phdcs.smu.edu.sg

Archan Misra
Singapore Management University
Singapore
archanm@smu.edu.sg

ABSTRACT

As RGB camera resolutions and frame-rates improve, their increased energy requirements make it challenging to deploy fast, efficient, and low-power applications on edge devices. Newer classes of sensors, such as the biologically inspired neuromorphic event-based camera, capture only changes in light intensity *per-pixel* to achieve operational superiority in sensing latency ($O(\mu s)$), energy consumption ($O(mW)$), high dynamic range ($140dB$), and task accuracy such as in object tracking, over traditional RGB camera streams. However, highly dynamic scenes can yield an event rate of up to 12MEvents/second, the processing of which could overwhelm resource-constrained edge devices. Efficient processing of high volumes of event data is crucial for ultra-fast machine vision on edge devices. In this poster, we present a profiler that processes simulated event streams from RGB videos into 6 variants of framed representations for DNN inference on an NVIDIA Jetson Orin AGX, a representative edge device. The profiler evaluates the trade-offs between the volume of events evaluated, the quality of the processed event representation, and processing time to present the design choices available to an edge-scale event camera-based application observing the same RGB scenes. We believe that this analysis opens up the exploration of novel system designs for real-time low-power event vision on edge devices.

CCS CONCEPTS

• **Computer systems organization → Embedded and cyber-physical systems**.

KEYWORDS

Edge AI, Machine Perception, Event Camera

ACM Reference Format:
Ila Gokarn and Archan Misra. 2024. Poster: Profiling Event Vision Processing on Edge Devices. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3643832.3661415

1 INTRODUCTION

Accurate and efficient visual perception on resource constrained edge devices is crucial to the realization of applications such as autonomous navigation on next-generation low-power edge platforms. Recent advancements in camera quality or resolution pose increased energy consumption during operation, creating a bottleneck to low-power efficient sensing on edge devices. In comparison, biologically-inspired neuromorphic event cameras mimic the human retina to provide extremely low power ($\sim 10 - 30mW$), highly reactive $O(\mu s)$ sensing capabilities, higher dynamic range up to $140dB$, showing competitive performance in DNN task accuracy against RGB cameras [8]. Neuromorphic event vision cameras (hereafter referred to as event cameras) move away from the CMOS sensor and concept of a “frame" to capture delta changes in light intensity (both positive and negative changes/polarity) incident on every pixel asynchronously, illustrated in Figure 1(a). As such, “images" can be synthesized/aggregated from a continuous stream of events that are reported as a tuple of ($pixel_x$, $pixel_y$, timestamp, polarity). These “images" or framed representations of the event stream can then be used as inputs to an off-the-shelf DNN, leveraging decades of computer vision research on RGB frames. However, highly dynamic scenes (for e.g., if the camera is in motion or observing many fast-moving objects) can yield a high volume of events ($O(MHz)$), which can quickly overwhelm the CPU on a resource-constrained edge device as it cycles through all the events to create a framed representation for DNN inference. Prior works on event vision disregard the event processing costs and either (i) focus on offline applications which post-process large volumes of event data for analysis, (ii) assume the deployment of a server-class GPU, or (iii) characterise event processing as an offline task, leaving real-time event processing at the edge an open problem.

In this work, we are the first to characterize the cost of event sensor processing on edge devices. We describe the design of our profiler and the trade-offs observed between the volume of events evaluated, pre-processing techniques used, quality of event representation generated, and processing time incurred, to derive the pareto optimal setting for event processing on the NVIDIA Jetson Orin AGX [7], a representative edge device. We present the profiler's operation over the CityFlow AI [2] person detection dataset.

2 PROFILER OVERVIEW AND RESULTS

The profiler comprises of 4 stages, described below.
1. Simulating Events: The profiler first synthesizes event streams from RGB videos using the v2e (video to events) software tool [3], aggregating events at a user-defined temporal resolution (default, t = 10ms or 100 FPS) to evaluate the expected volume of events if an event camera were to replace the RGB camera. The profiler's goal is to identify the best {framed representation, pre-processing technique} combination for the given dataset which provides the best quality of framed representation of the events before the next window of events is aggregated i.e. within t of 10ms.

MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06.
https://doi.org/10.1145/3643832.3661415

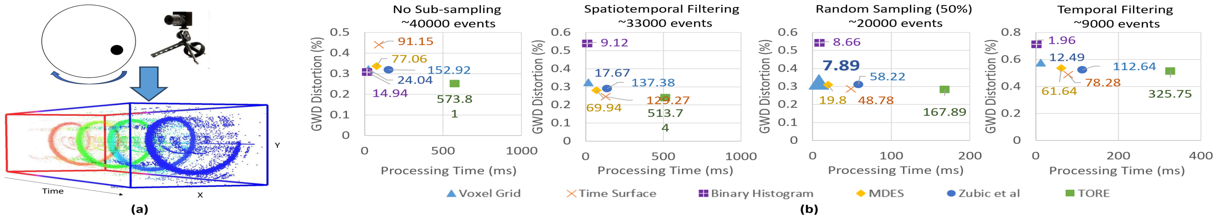

Figure 1: (a) Event stream generated when observing a marker-equipped rotating disk (b) Distortion rate (lower, better) vs processing time (lower, better) of ~40000 events accumulated in t = 10ms

2. Applying Pre-processing Techniques: The profiler then evaluates sampling techniques to reduce the volume of events and the resulting processing time: (i) No sampling (ii) Spatiotemporal filtering (drop isolated events in a 1 pixel radius over 1ms window) (iii) Temporal downsampling (drop every 2nd event per-pixel) (iv) Random sampling (drop events with $p = 0.5$ probability).

3. Creating Framed Representations: Next, the profiler loads the events in each time bin into memory to convert the events into one of the following event representations: (i) Voxel Grid (ii) Binary Histogram (iii) Time Surface [5] (iv) Mixed Density Event Stack (MDES) [6], (v) 12-channel representation [9], and (vi) Time Ordered Recent Events (TORE) [1].

4. Calculating Quality of Generated Representation: Lastly, the profiler leverages the Gromov-Wasserstein Discrepancy (GWD) as a metric for comparing the quality of event representations efficiently by measuring the distortion arising from the conversion of raw events to the framed representations. The GWD metric calculates the similarity between event and feature pairs during the construction of an event representation, with a lower GWD score indicating lesser distortion (or equivalent, better representation quality and preservation of events) and better DNN accuracy [9].

2.0.1 Evaluation Results. In our experiments we utilize the the CityFlow AI dataset [2] (1920 × 1080 resolution videos captured at 30FPS) for a person detection application. In lieu of comparing multiple datasets that generate different event volumes, we utilize the *same* dataset and vary the sensor resolution settings to control the average number of events accumulated in a single time bin, simulating for both (i) sudden/unexpected bursts of high volume events and (ii) different choices of event camera resolutions. We simulate events with resolution settings (120×90, 346×260 (the resolution of DVS346 [4], a widely deployed event camera), 480 × 270, 512 × 290, 640×480) to generate (10000,..., 50000) events on average in a single time bin of t = 10ms to understand the processing latency on the Jetson AGX Orin over different event volumes. Table 1 describes a linear relationship between event volume and processing time, with Voxel Grids and Binary Histograms processing up to 20000 events within the processing deadline t of 10ms. This indicates that in general, if we are able to pre-process high volumes of event data down to ~20000 events, we can achieve fast non-blocking event processing on a Jetson AGX Orin. Figure 1 describes the trade-off between distortion and processing time with pre-processing techniques applied to an event stream that generates ~40000 events on average in a 10ms window. Nuanced event representations (e.g. MDES and TORE) achieve a lower distortion rate ≤ 30% but suffer 30×−50× higher processing times than the required 10ms. Random sampling gives the most amount of control over the volume of events filtered without suffering much distortion compared to the original event stream. Temporal filtering on the other hand suffers the most distortion with no relative gains in processing time. Finally, the profiler determines that for the CityFlow AI dataset for person detection [2], Voxel Grids can achieve both a lower distortion rate of 0.31% and a fast processing time of 7.89ms when random sampling is applied to reduce the volume of events to ~20000 events.

Framed Representation	Processing Time of Event Volume (ms)				
	10000	20000	30000	40000	50000
Voxel Grid	3.4	6.35	11.36	24.04	38.97
Time Surface [5]	34.82	45.63	52.49	91.15	128.4
Binary Histogram	3.17	4.38	6.81	14.94	21.87
MDES [6]	8.63	16.75	24.11	77.06	99.52
Zubic et a [9]l	28.57	46.23	61.72	152.92	196.67
TORE [1]	99.51	173.2	202.88	573.81	884.01

Table 1: Processing time vs event volumes in a 10ms window

ACKNOWLEDGMENTS

This work was supported by National Research Foundation, Prime Minister's Office, Singapore under both its NRF Investigatorship grant (NRF-NRFI05-2019-0007), and its Campus for Research Excellence and Technological Enterprise (CREATE) program. The Mens, Manus, and Machina (M3S) is an interdisciplinary research group (IRG) of the Singapore MIT Alliance for Research and Technology (SMART) centre. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

[1] R. W. Baldwin, R. Liu, M. Almatrafi, V. Asari, and K. Hirakawa. 2022. Time-ordered recent event (TORE) volumes for event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2022), 2519–2532.

[2] Milind Naphade et. al. 2023. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*

[3] Y Hu, S C Liu, and T Delbruck. 2021. v2e: From Video Frames to Realistic DVS Events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE.

[4] Inivation. 2024. Inivation DVS346. https://tinyurl.com/4bsv5acx

[5] Xavier Lagorce, Garrick Orchard, Francesco Gallupi, Bertram E. Shi, and Ryad Benosman. 2017. HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE T-PAMI* 39, 7 (2017), 1346–1359.

[6] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. 2022. Stereo depth from events cameras: Concentrate and focus on the future. In *Conference of Computer Vision and Pattern Recognition (CVPR)*. 6114–6123.

[7] NVIDIA. 2024. Jetson Orin AGX. https://tinyurl.com/4yc8ny93

[8] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. 2020. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems* 33 (2020), 16639–16652.

[9] Nikola Zubić, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. 2023. From chaos comes order: Ordering event representations for object recognition and detection. In *Proceedings of IEEE/CVF ICCV*. 12846–12856.