https://doi.org/10.31256/Ik8Mv1B

Event-based Spatio-temporal down-sampling

Anindya Ghosh
Dept. of Informatics
University of Sussex
Brighton, UK
Email: anindya.ghosh@sussex.ac.uk

Thomas Nowotny
Dept. of Informatics
University of Sussex
Brighton, UK
Email: t.nowotny@sussex.ac.uk

James Knight
Dept. of Informatics
University of Sussex
Brighton, UK
Email: j.c.knight@sussex.ac.uk

***Abstract*—As vision sensors for autonomous systems, event based cameras provide numerous benefits over conventional cameras including higher dynamic range and temporal resolution as well as lower bandwidth and power requirements. However, while down-sampling is regularly used in standard computer vision, there are no reliable technique to do this for event data, resulting in a bottleneck for event based computer vision systems. Here we propose two novel methods for down-sampling event data and compare them against a previously used method.**

***Index Terms*—sensors, event cameras, bio-inspired vision**

I. Introduction

Event cameras [1] are bio-inspired, neuromorphic sensors that integrate per-pixel luminance changes and emit events whenever these exceed a certain threshold; i.e., unlike conventional cameras that capture frames at a fixed rate, these devices asynchronously emit events (x_i, y_i, p_i, t_i) where x_i and y_i are the coordinates of the event, t_i is the time it occurred at and $p_i \in \{\mathrm{ON}, \mathrm{OFF}\}$ is its 'polarity' indicating whether this event represents an increase or decrease in luminance.

Event cameras offer higher dynamic range, microsecond temporal resolution and higher data throughput compared to their conventional counterparts. However, in event cameras, only contrast changes are recorded so any luminance information that might be pertinent to later post-processing – such as down-sampling – is unavailable.

In computer vision, when additional image resolution no longer improves performance, down-sampling input images can be a simple way of reducing memory and computational costs. There are a plethora of algorithms for down-sampling frame-based data; and although the resolution of event-based cameras is on the rise, the same scope of algorithms does not exist for the latter. This results in issues such as shadows and reflections in complex stimuli being largely ignored and limits reasonable down-sampling to fairly simple stimuli. Some spike-based works circumvent this issue by simply generating frames from events, down-sampling the frames and subtracting them from each other to recover ON and OFF polarity events at lower resolution [2]. However, this approach loses many of the advantages of event cameras and issues such as latency and motion blur are reintroduced. Other works that have used spatio-temporal down-sampling of event data streams, either implement the technique on simple stimuli (devoid of shadows or reflections), or the down-sampling scale in the context of the stimuli is low enough to render the particular down-sampling technique inconsequential [3].

The aim through this work, therefore, is to obtain a down-sampled event stream that mimics the output of a lower resolution event-based camera, irrespective of the complexity of the stimuli. Here we present two down-sampling methods that provide down-sampled event data that better approximates a lower resolution event-based camera than previous methods while still requiring minimal computational overhead.

II. Methods

A. Simulated Events

As event cameras with arbitrary resolutions are not available, we generated simulated events in order to test our methods. We used ESIM [4] to generate events from high resolution (640 × 360) and down-sampled (64 × 36) video frames. To enable a better approximation of the high temporal resolution of event-based cameras we used 1200fps footage of a cheetah. Fig. 1 shows example input frames and output events (accumulated over 25 ms for visualisation). Note that the camera is centred on the cheetah so, in Fig. 1B, the grass moving beneath it is the largest source of events. However, this high-frequency detail is almost entirely missing from Fig. 1D.

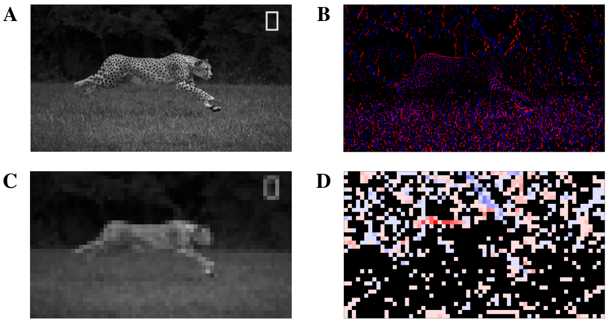

Fig. 1: **(A)** High-resolution frame **(B)** Events, generated from A with ESIM. Red, blue and black pixels signify ON, OFF and no spikes respectively. **(C)** Down-sampled frame **(D)** Events generated from C with ESIM. Darker shades signify higher number of spikes of respective polarity in corresponding pixel bin.

B. Down-sampling Techniques

Given events (x_i, y_i, p_i, t_i) – we wish to *spatially* down-sample x_i and y_i by s_x and s_y respectively; and *temporally*

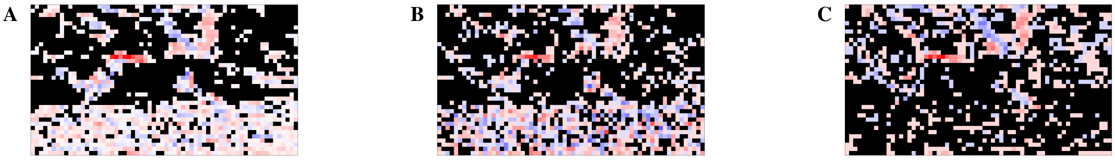

Fig. 2: Visual illustration of the three down-sampling techniques. **(A)** Naive down-sampling. **(B)** Binning down-sampling. **(C)** IF down-sampling.

down-sample t_i by s_t – resulting in down-sampled events $(\lfloor x_i \times s_x \rfloor, \lfloor y_i \times s_y \rfloor, p_i, \lfloor t_i \times s_t \rfloor)$.

1) Naive: The naive approach [5] is to simply spatially down-sample each event individually. However, this does not reduce the number of events, results in unrealistically high firing rates for individual down-sampled pixels and is sensitive to high-frequency noise which would not be visible to an actual low resolution event camera.

2) Binning: Since event-based cameras have a very high temporal resolution, multiple ON and OFF polarity events can occur in the same pixel within a down-sampled timestep. This method attempts to extract better down-sampled events by taking into account this fine temporal structure. This is done by first subtracting the total number of OFF events from the total number of ON events in each down-sampled pixel during an intermediate timestep of $33\,\mu\text{s}$. Next, the number of intermediate timesteps where this sum is positive and the number where it is negative are summed across the full down-sampled timestep. Finally, if the number of intermediate timesteps with a positive sum is larger than it was during the previous down-sampled timestep, a ON event is emitted and, if the number of intermediate timesteps with a negative sum is larger than it was during the previous timestep, an OFF event is emitted.

3) Integrate-and-fire: Within each down-sampled pixel, this method also accumulates events. However, inspired by integrate-and-fire neurons, this accumulation occurs *across* down-sampled timesteps. Each down-sampled pixel has a separate accumulator for each event polarity and when these reach a threshold, an event of the correct polarity is emitted and the counter is reset to zero.

III. Results

Fig. 2 shows down-sampled images from each of the three techniques after having applied the techniques to a full-resolution video.

A. Qualitative Analysis

Fig. 1D shows that few events were generated by the moving grass due to the frame-based down-sampling removing such high-frequency texture. While this is clearly not the case with the naively down-sampled events shown in Fig. 2A, the events generated by our improved algorithms – shown in Figs. 2B and C – both clearly illustrate this effect.

B. Quantitative Analysis

Here we compare the spike trains generated directly from 640×360 frames down-sampled to 64×36 using the OpenCV INTER_AREA scale filter algorithm, to those down-sampled using one of the event-based methods we describe above. We compare the spikes emitted by each approach in a $25\,\text{ms}$ 'timestep' using the structural similarity index measure (SSIM) which aims to measure the perceived change in the structural information of the image [6]. We then calculated the mean-squared-error (MSE) across all timesteps in our $800\,\text{ms}$ video. While the naive approach only achieved an SSIM of 0.136, the binning and integrate-and-fire approaches achieved an SSIM of 0.156 and 0.353 respectively, representing a significant improvement in event-based down-sampling performance. Furthermore, there is also a significant reduction in the total number of spikes generated over the course of the $800\,\text{ms}$ video with the naive approach generating 84,188 spikes and the binning and integrate-and-fire approaches producing 44,586 and 26,352 spikes respectively – significantly reducing the processing costs of the down-sampled events. In comparison, the benchmark spike train generated by ESIM has 26,300 spikes.

IV. Conclusions

Two novel methods were presented against a method that is generally used for spatio-temporal down-sampling of event-based data. The two methods outperform the incumbent technique. This was tested by comparing the event data streams of the two methods against a benchmark down-sampled simulated event data stream obtained using ESIM.

V. Acknowledgments

AG is funded by the Leverhulme Trust. JK was funded by the EPSRC (grant EP/V052241/1).

References

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[2] M. Blanchard, F. C. Rind, and P. F. Verschure, "Collision avoidance using a model of the locust lgmd neuron," *Robotics and Autonomous Systems*, vol. 30, no. 1-2, pp. 17–38, 2000.

[3] L. Salt, G. Indiveri, and Y. Sandamirskaya, "Obstacle avoidance with lgmd neuron: towards a neuromorphic uav implementation," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2017, pp. 1–4.

[4] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: an open event camera simulator," *Conf. on Robotics Learning (CoRL)*, Oct. 2018.

[5] G. Cohen, S. Afshar, G. Orchard, J. Tapson, R. Benosman, and A. van Schaik, "Spatial and temporal downsampling in event-based visual classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 5030–5044, 2018.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.