DVS-OUTLAB: A Neuromorphic Event-Based Long Time Monitoring Dataset for Real-World Outdoor Scenarios

Tobias Bolten, Regina Pohle-Fröhlich
Hochschule Niederrhein
Institute of Pattern Recognition
Krefeld, Germany
{tobias.bolten, regina.pohle}@hs-niederrhein.de

Klaus D. Tönnies
University of Magdeburg
Department of Simulation and Graphics
Magdeburg, Germany
klaus@isg.cs.uni-magdeburg.de

Abstract

Neuromorphic vision sensors are biologically inspired devices which differ fundamentally from well known frame-based sensors. Even though developments in this research area are increasing, applications that rely entirely on event cameras are still relatively rare. This becomes particularly clear when considering real outdoor scenarios apart from laboratory conditions.

One obstacle to the development of event-based vision applications in this context may be the lack of labeled datasets for algorithm development and evaluation. Therefore we describe a recording setting of a DVS-based long time monitoring of an urban public area and provide labeled DVS data that also contain effects of environmental outdoor influences recorded in this process. We also describe the processing chain used for label generation, as well as results from a performed denoising benchmark utilizing various spatio-temporal event stream filters.

The dataset contains almost 7 hours of real world outdoor event-data with ≈47k labeled regions of interest and can be downloaded at http://dnt.kr.hsnr.de/DVS-OUTLAB/

1. Introduction

The proposed "DVS-OUTLAB"-dataset was created as a part of a higher-level project, which aims to improve the urban planning of public open spaces by including the user behavior into this planning step. Aiming at this goal long-term measurements of a publicly accessible outdoor area were carried out over a period of several months. We present the dataset and the description of the technical setup that was used to carry out these observations by utilizing three Dynamic Vision Sensors (DVS).

Dynamic Vision Sensors are the result of the ongoing research in the field of neuromorphic engineering. They are bio-inspired vision sensors with the operating paradigm that all pixels work independently and asynchronously from each other. Each pixel can trigger an output based on detected local brightness changes exceeding a defined threshold [11].

This results in a sparse output data stream of activated pixels at a variable data rate. Each of these pixel activations is called an "event" and carries information about (a) the local (x, y)-coordinate within the sensor-array of the activated pixel, (b) the timestamp of activation for this event and (c) the polarity of the event which indicates the direction of the brightness change. The operating scheme of a DVS leads to a reduced data redundancy, a higher temporal resolution, lower power consumption and a higher dynamic range compared to classical image sensors. These properties are especially beneficial in outdoor recording settings. A further advantage of the DVS technology is that it offers the opportunity to carry out the monitoring under low privacy regulations. In comparison to classical frame based vision, no gray or color values need to be processed by any software logic.

However, while in the context of frame based video monitoring and surveillance several annotated datasets for tasks like action or anomaly detection (e.g. [21, 27, 31]) or traffic flow analysis (e.g. [25, 35]) exist, there is currently a lack of available event based annotated datasets.

This shortcoming is even more apparent considering outdoor usage scenarios as in contrast to laboratory conditions the recorded event stream also contains artefacts from environmental influences. Summarizing our main contributions we therefore provide

- a DVS based dataset containing almost 7 hours of raw event data in the context of static long-term monitoring of a public outdoor area
- the description of a semi-automatical labeling chain and the resulting selection of almost 50k labeled regions of interest

- as well as benchmarking results of spatio-temporal filters to address the challenge of denoising DVS event streams.

The rest of this paper is structured as follows. Section 2 briefly summarizes the requirements for data recordings in the context of the performed measurement and provides an overview of already existing datasets and approaches. A description of the technical setup used for data collection onsite follows in Section 3. Section 4 introduces the labeling processing chain as well as statistics over the provided data. In Section 5 the denoising benchmarking is presented. Finally, a short summary is given in Section 6.

2. Use-case scenario

2.1. Requirements

In the context of the performed stationary long time monitoring (details are following in Section 3) a dataset is required that

(a) contains only limited ego-motion due to the usage of fixed mounted sensors.

(b) includes challenging illuminations, sensor noise and environmental interferences (compare to Figure 2) through the outdoor measurement.

(c) provides suitable label annotations and classes (compare to given labels in Table 2) for the monitoring of public open places.

In machine learning, especially when using deep learning techniques, often large datasets are necessary. Only these datasets allow a reliable and comparable development of specific algorithms or complete processing pipelines. Therefore it is also required that

(d) the database is large enough to support training in a deep learning approach, e.g., for object detection or tracking.

2.2. Existing DVS-Datasets

Frame based DVS-Stream simulation

Early work often used well known frame-based datasets by simply recording a computer display with a DVS. In this case events were triggered by flashing or moving the image on the screen or through the simulation of small eye movements, so-called saccades, by moving the sensor in front of the screen. Examples for these kinds of datasets are several converted versions of the MNIST dataset (e.g. N-MNIST [28], MNIST-DVS [33]) or the N-Caltech101 [28] dataset. Besides the missing applicability of these datasets with respect to the described monitoring scenario, the time-continuous aspect of the DVS event stream is also missing by converting individual image frames.

Frame-based video datasets have also been converted [19] to address this limitation. However, it is difficult to use in the sense of long-term monitoring, because only very short sequences of very different sceneries were considered.

Furthermore, this kind of conversion by recording scenes displayed on a screen does not fully include realistic DVS-characteristics. E.g., the time resolution of events is limited by the frame rate of the shown video material and the refresh rate of the used screen. Therefore a computationally realistic simulation of DVS output from frame based material is an active field of research [6, 12, 32]. The "v2e"-framework introduces a DVS pixel model including temporal noise, leak events, finite bandwidth as well as a Gaussian threshold distribution into the simulation process [10]. To overcome the time resolution limitation of classical frame based video input material this framework creates interpolated frames by applying a synthetic slow motion deep learning approach. Although this seems to be an interesting and promising path to accelerate the DVS-based work by re-using the effort already invested in existing frame-based datasets, this approach does not fit very well to the needs in terms of a long-term monitoring scenario.

Frame-based datasets in this context are usually recorded by closed CCTV video surveillance systems operating at low image resolutions and/or frame-rates. Examples are the datasets "ADOC" [31] with 3 fps or "ICVL" [21] with 15 fps, each containing several hours of footage. The upper limit are usually 30 fps recordings (like "VIRAT" [27]), which would still require a high up-sapling factor to achieve typical DVS time resolution of a few milliseconds within the simulation. In addition, the computation time required to simulate DVS-data leads to another practical limitation. The authors of the v2e-framework estimate that the simulation runs between 20 to 100 times slower than realtime [10]. Therefore, the conversion of large datasets with multiple hours of footage is quite challenging.

Native DVS-Datasets

Most of the existing published neuromorphic datasets are composed of temporally short sequences of specific actions or patterns recorded well aligned in mostly clean laboratory environments. The "POKER-DVS"-dataset [33] (sequences of fast flipping poker cards) or "DVS128 Gesture" [2], "SL-ANIMALS-DVS" [36] (sequences of hand/sign language gestures) are examples.

Available, more general datasets, recorded under real world conditions relate to autonomous driving [9, 18, 30, 34]. However, due to the movement of the vehicle, these recordings contain a high proportion of egomotion which makes them difficult to apply in the required monitoring application context.

Currently there exists just one dataset [26] covering the

related task of pedestrian detection. However, this part of the dataset consists only of 12 recordings with an average length of 30 seconds. In summary, no published DVS dataset currently fulfills the requirements resulting from the long-term monitoring of public urban areas.

3. Dataset Recording setup: Living-Lab

To allow the inclusion of users and their behavior into urban planning processes a public children playground was considered as a living-lab concept. The "DVS-OUTLAB" dataset consists of selected recordings from multiple Dynamic Vision Sensors which carried out a long-time monitoring of this playground.

Technical setup

Figure 1a depicts a schematic plan of the monitored playground. The area under surveillance has an approximate size of 2800 square meters and is observed by three fixed mounted sensors. Each of the Dynamic Vision Sensors is mounted in a weatherproof enclosure on a mast at a height of approx. 6 meters with a pitch angle of about 25 degree to ground (see Figure 1b). The positioning of the sensors has been chosen in such a manner that blind spots given due to the terrain characteristics (trees, bushes, mounds) are minimized. The three sensor positions are connected by underground cables to a central point (see Figure 1c) for sensor data acquisition. The complete system is powered by an independent solar-based energy storage.

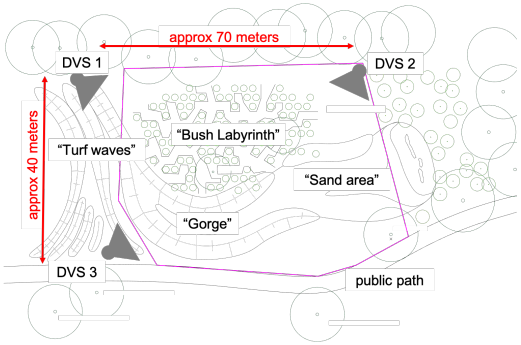
Used Sensor

In this measurement setup we used three models of the CeleX-4 DVS-Sensors [15], each equipped with an 8mm wide-angle lens[1]. The CeleX-4 sensor is the fourth generation of a neuromorphic vision sensor series [8, 14, 20, 7] developed and distributed by Celepixel Technology[2]. This sensor offers a high-speed event output with 200Meps (events per second), a high dynamic range and a spatial resolution of 768×640 pixels [15, 11].

The sensor array of the CeleX-4 DVS is logically divided into 5 read-out blocks, each consisting of 128 rows. According to the feedback from the manufacturer the first block containing the upper 128 pixel rows is read out more frequently than the others. This leads to different event-triggering frequencies within the pixel array and has also implications to the other blocks due to the performed off-sensor event timestamping. For further notes on this sensor behavior compare with issue 8 in [24]. To avoid these problems, we have completely disabled this block in our recordings, resulting in a remaining resolution of 768×512 pixels.

[1] Computar V0814-MP, f=8mm, F1.4, 1", C-Mount

[2] https://www.celepixel.com/



(a) Schematic plan of the measuring area and DVS positioning (each sensor is equipped with an 8mm focal length wide angle lens mounted in 6 meter height at the head of a mast)

(b) Sensor mount and case

(c) Central data acquisition and processing point

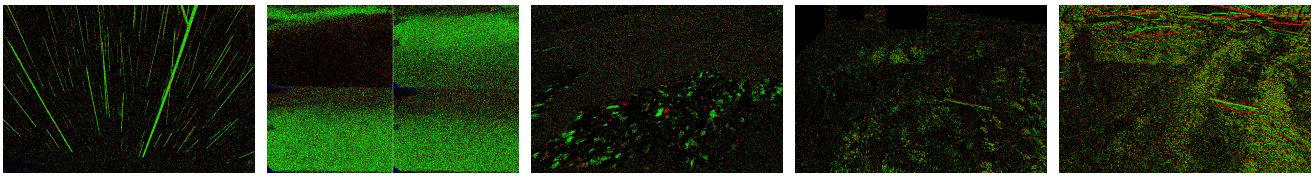
(d) FoV from DVS1 as greyscale image for clarification with typical object heights for an adult person depending on the distance to the sensor

Figure 1: Outdoor recording area and sensor mount setup

To illustrate the recorded scene, Figure 1d shows the acquired field-of-view of one of the sensors as a gray-scale image.

Privacy focus

The described living-lab measurement setup is located in Germany. Due to very strict data protection laws [3, 5, 13] and a high expectancy of the potential users for privacy (especially in the considered use-case of a children playground), it was not possible to use additional sensors such



(a) Rain (b) Global lighting change by clouds (c) Shadows caused by tree crowns (d) Movement of tall grass due to wind (e) Ego-Motion caused by minimal mast movements

Figure 2: Visualization of environmental influences caused by the outdoor scenario. (Each image pixel encodes the polarity of the last occurred event within a time window of 60ms. A brightness increase is displayed in green and a decrease in red. Short video sequences are also provided at the dataset webpage for a better impression.)

as classical RGB-frame based CCTV systems besides the DVS.

At this point using DVS-technology falls under significantly lower regulations due to inherent sensor properties. No gray or color values have to be processed by the application logic during the subsequent analysis of the acquired data stream.

4. Provided Database

The presented dataset includes data from two semantical different recording scenarios of the same area. The first recording session contains staged scenes, which were subsequently semi-automatically tagged with event-wise semantic class labels. The data set from the second recording session also includes artefacts from environmental influences.

In the Subsection 4.1 some properties which pose a challenge in subsequent semantical processing steps are briefly presented. The label generation process, as well as statistics on the published data, are described in the Subsections 4.2 and 4.3.

4.1. Challenges

Due to the fixed mounted sensor measurement setup the dataset contains just minor parts of ego-motion background cluttering. Still, the realistic outdoor setting produces data properties, that make the development of high level computer vision solutions challenging.

Background noise: The output of currently available Dynamic Vision Sensors contains noise, caused for example by junction leakage currents and thermal noise. In [11], Gallego et al. give an overview of the specifications of different DVS models, including a consideration for the stationary noise behavior. From this comparison it can be seen that the event stream of the CeleX4-sensor contains a relatively high proportion of noise.

Object sizes: The low resolution, compared to state of the art frame-based systems, of the CeleX-4 DVS sensor in combination with the used wide-angle lens and the size of the monitored area leads to small object sizes. Figure 1d illustrates this by means of apparent object sizes of an adult person, which varies between 17px and 110px depending on the distance to the sensor.

Environmental interferences: Especially in outdoor usage scenarios the Dynamic Vision Sensors tends to capture different kinds of environmental interferences due to its high sensitivity and temporal resolution. These interferences are mostly not taken into account in other datasets.

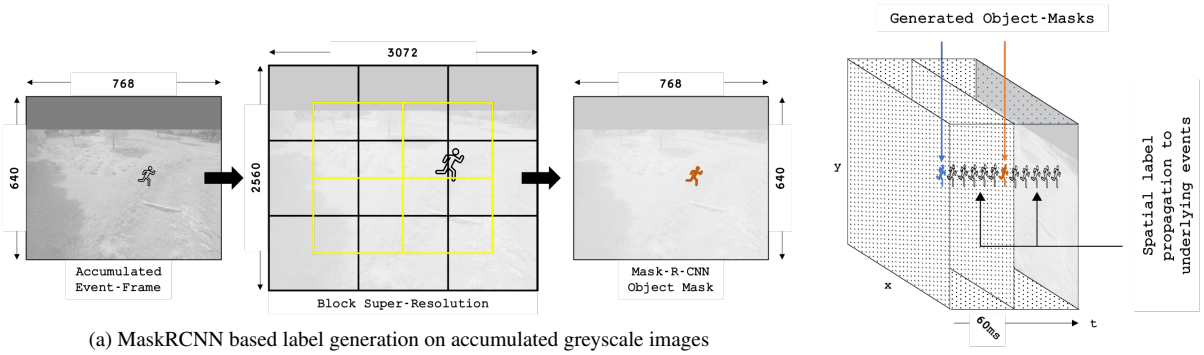
Figure 2 shows an example of some of these disturbances such as rain or shadows, where the DVS event stream data is displayed as event polarity image.

4.2. Staged scenes and semi-automatic labeling

To be able to generate a large annotated dataset, we restaged typical use scenarios at the site. During the recordings of these scenes, we were also able to temporarily acquire and store the brightness values provided by the CeleX sensor stream. This was only possible by explicit consent of the observed persons due to the privacy concerns as mentioned in Section 3, while the area was closed for public usage. In this way, several hours of material could be recorded that included activity, which was then semi-automatically augmented with object label annotations.

Label Generation

In the context of this special recording setup it was possible to calculate grayscale images based on the brightness values provided per event directly by the CeleX data stream. Using these images, generated at a time interval of 60ms, well known frame-based CNN methods were utilized to generate object label proposals. For this purpose, we used an implementation [1] of the Mask-R-CNN object detector [17] which was pre-trained on the COCO database [23].



(a) MaskRCNN based label generation on accumulated greyscale images

(b) Label propagation from object masks to event stream

Figure 3: Processing step visualization for label generation and propagation on stages dataset scenes

Sensor	#TimeWindow recorded	#TimeWindow containing Label
DVS1	131492	98853
DVS2	131488	52520
DVS3	131488	76337
Total	394468 ≈6h 30min	227710 ≈4h

Table 1: Staged scenes: number of total recordings and automatic label suggestions (TimeWindow $\hat{=}$ 60ms)

We first applied a deep learning super-resolution network [22] to scale the input images by the factor of 4. This should mitigate effects from the low image resolution of acquired grey scale images on small objects. The scaled images were divided into 13 blocks so that the block-width corresponds to the expected input size of the used Mask-R-CNN network. For each block, the result of the Mask-R-CNN inference was calculated and combined into a single mask image of the size of the original image. This procedure is depicted in Figure 3a.

Based on this object mask, the generated labels were propagated to the events of the corresponding 60ms time segment of the initial event stream (see Figure 3b). In this step, objects were filtered out that did not move in the corresponding time windows, since they are visible in the greyscale images but do not have a significant amount of associated events in the DVS stream.

4.3. Dataset Statistics

The total length and number of automatically generated label-proposals by the Mask-R-CNN pipeline are summarized in Table 1. Since almost all staged scenes have a person involved, this results in a very large class imbalance in the class appearance frequency. For this reason, we propose to use only a sub-selection of the available data.

In the process of sub-selection, attention was paid to the following two aspects. First to select an equal number of examples per class and sensor row-block, if it was available in the original data. The goal of this row-block guided sub-selection is to equally include all different object sizes resulting from the different object to sensor distances included per block. Secondly, care was taken to ensure the label quality of the Mask-R-CNN predictions for the selected data. The labeling was manually controlled by a human and only satisfying predictions were added into the final selection.

The number of resulting labeled event time windows is given in Table 2. Each time window contains labels within an 192×128px region of interest when a selected object is included in this region. Examples for selected regions are given in Figure 4. For a visual impression of the achieved label quality a comprehensive overview of label visualizations is available on the dataset webpage.

We also propose a 70/15/15% split that can be used to train, validate and test further work and applications. The distribution across the sensor row-blocks has also been taken into account by performing this split.

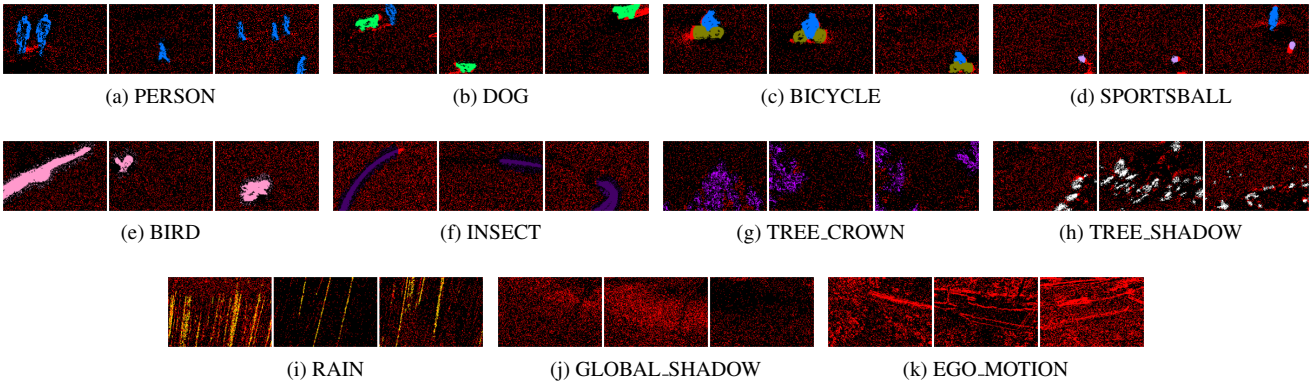
5. Spatio-temporal Event-Filter Comparison

As already mentioned in Section 4.1 currently available Dynamic Vision Sensors suffer from noise. Therefore, denoising of DVS event streams is an active research topic in which we will present the first application of our database.

In the context of long-term monitoring, a fast computational and cost-effective approach to noise filtering is needed. For our dataset, nearly 42 billion events were captured over a total of almost 7 hours of recordings. With a storage requirement of 80 bits per event as it is implemented

	Class-Label	#TimeWindow containing Label	#TimeWindows containing Label in Row 1	Row 2	Row 3	Row 4
Objects of interest	PERSON	7399	2833	2678	1255	633
	DOG	709	351	259	54	45
	BICYCLE	4378	2023	1834	478	43
	SPORTSBALL	500	147	244	74	35
Environmental interferences	BIRD	3807	825	1353	1173	456
	INSECT	5939	842	1329	1532	2236
	TREE_CROWN	6375	1731	2511	1576	557
	TREE_SHADOW	6901	50	1660	2387	2804
	RAIN	7052	1776	1776	1750	1750
	GLOBAL_SHADOW	4800	1200	1200	1200	1200
	EGO_MOTION	4800	1200	1200	1200	1200

Table 2: Performed sub-selection from all labeled event data. Each row represents a 128 pixel sensor block on the Y-axis, starting to count at the top. (TimeWindow $\hat{=}$ 60ms)



(a) PERSON (b) DOG (c) BICYCLE (d) SPORTSBALL

(e) BIRD (f) INSECT (g) TREE_CROWN (h) TREE_SHADOW

(i) RAIN (j) GLOBAL_SHADOW (k) EGO_MOTION

Figure 4: Example snippets from provided database (more examples are provided on the dataset webpage)

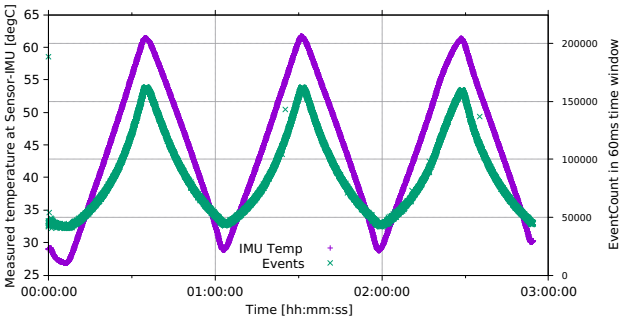


Figure 5: CeleX4 sensor temperature correlated noise behavior (recorded scene is unchanged in temperature controlled environment)

by the CeleX SDK [24], this results in approximately 390 gigabytes of data. When measuring with three sensors for 10 hours per day as in our real application scenario, denoising for data reduction is unavoidable. The Figure 5 illustrates that the amount of sensor background noise also increases significantly with higher ambient temperature of the sensor. This behavior was also evident in our recorded data, because in summer temperatures of over 50 degrees were measured inside the case. Therefore, reducing the number of events in the data by efficiently eliminating event noise represents the first processing step in the analysis of our data. Due to the limited computational and power capacities on site and the need to process three DVS streams in parallel in real time at a central location, as mentioned in the description of the technical setup above, the use of simple spatio-temporal event filters seems to be a suitable solution.

Subsequently we define different spatio-temporal filters and compare their results on the provided dataset. A filter is suitable for our application if it has the following properties:

- removing as many events as possible that were triggered by noise,

- eliminating as many events as possible that were generated by environmental influences and
- obtaining as many events as possible that were created by objects of interest.

5.1. Filter-Logic

The basic assumption of spatio-temporal filters is that events from real objects should occur spatially and temporally more often than noise events [11].

Neighborhood-Filter: For each event e spatially adjacent events are evaluated in its time windows tw_i and its preceding window tw_{i-1}. The number of populated spatial neighborhood cells that include at least one other event is counted. For this a 8-neighborhood is used for tw_i and a 4-neighborhood in the previous tw_{i-1}. The event e is discarded if less than thres_1 neighborhood cells for tw_i are populated or thres_2 cells for tw_{i-1}.

In the following experiments we set $\text{thres}_1 = 4$ and $\text{thres}_2 = 2$.

Time-Filter: For each event e it is checked whether or not there was another event at the same (x,y) position in the preceding x milliseconds. If no other event occurred in this timespan, the event e is considered as noise.

In the following experiments we calculated different results for the time threshold x=3ms, x=6.5ms and x=10ms.

SeqX-Filter: For each event e the spatial distances to a small number of directly preceding events in the acquired sensor stream is calculated [16]. If the smallest occurred distance is below a defined threshold, the event is kept. Further details can be found in [16].

In the following experiments we set the number of considered preceding events to 10 and the threshold σ = 0.01.

EDnCNN: It consists of three 3×3 convolutional layers (using ReLU, batch normalization and dropout) and two fully connected layers [4]. An Adam optimizer with a decay rate of 0.1 and a learning rate of 2E-4 is used for learning. For each event, the network makes a binary classification based on a feature vector generated from the spatial-temporal neighborhood.

5.2. Filter-Results

Comparison on provided data and scenario

We compare the performance that can be achieved using these different filters. The EDnCNN approach was trained for this purpose with our data.

Similar to [29] we calculate the **P**ercentage of **R**emaining **E**vents (PER) after filtering for each labeled 60ms time windows of the dataset. However, we additionally consider the effects of filtering for each individual object class as well:

$$\text{PRE}_c^f = \frac{\#\text{Events}_c^f}{\#\text{Events}_c^{total}} \cdot 100 \tag{1}$$

where f describes the used filter method, c the object class and #Events the number of corresponding events.

The selected labels of the dataset (see Section 4.3 and Table 2) are scattered over the complete duration of the dataset. However, for the filter calculation the entire temporal sequence (also of unlabeled events) were considered, but only labeled events were judged in the benchmarking. Figure 6 summarizes the distribution of per time window calculated PRE filter results considering all selected and labeled parts of this dataset.

It can be noted that by means of these spatio-temporal filters generally a significant reduction of background activity noise can be achieved. However, the proposed filters differ in their results with respect to individual object classes. The EDnCNN achieves good denoising results on background noise as well as on effects like global shadow and ego motion. Nonetheless this approach also preserves a relatively high proportions of other environmental interferences. In turn, the 'Neighborhood-Filter' achieves a higher denoising level on these environmental interferences classes, but also removes a slightly higher amount of other object class events. However, in total, the effects of environment influences can only be removed to a limited extent and must be adressed separately in subsequent analysis.

As the object class `PERSON` in Figure 7 shows, the average PRE result for each filter is individually comparable when considering the different sensor row-blocks (128px on Y-axis) separately. However, the variation within the results decreases with an increasing object size.

Comparison in other scenarios

In a second step, we evaluate the results on the DVSNOISE20 dataset[3], which is explicitly provided by the EDnCNN authors to evaluate event denoising algorithm performance against real sensor data. The data included in this dataset contains recordings with moderate to very heavy ego-motion of scenes with differently textured details. For comparison, we therefore examined only the recordings from the DVSNOISE dataset, which showed a low ego-motion/event count and most closely matches to the application in the described Living-Lab.

[3] https://sites.google.com/a/udayton.edu/issl/software/dataset

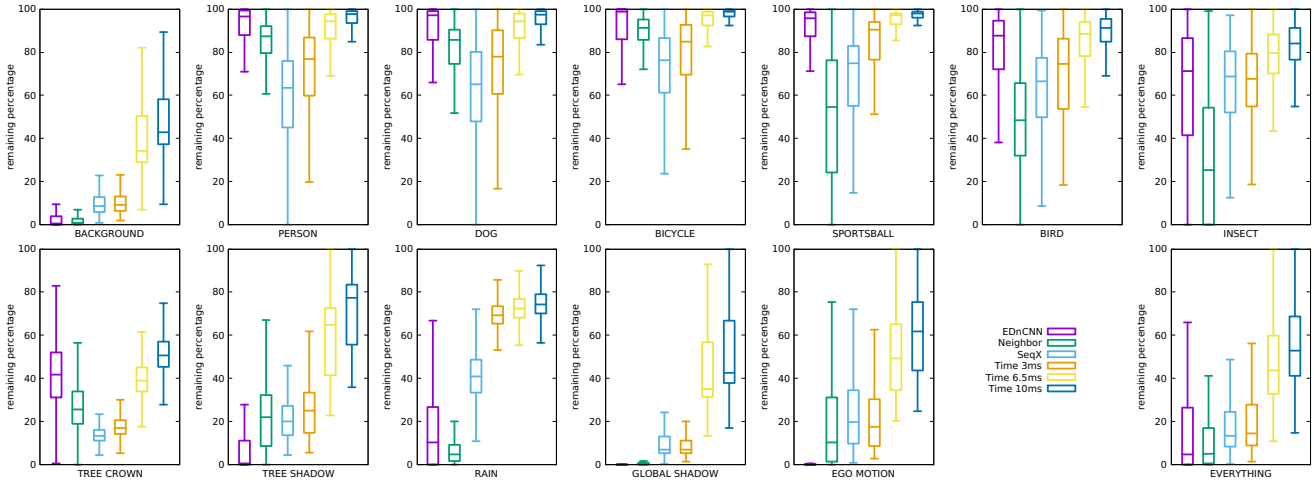


Figure 6: PRE results from spatiotemporal filters on the provided DVS-OUTLAB dataset

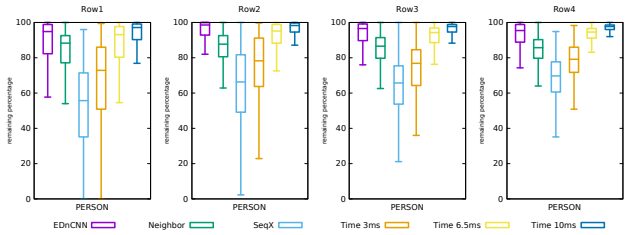


Figure 7: PRE filter results for class `PERSON` separated by sensor row blocks

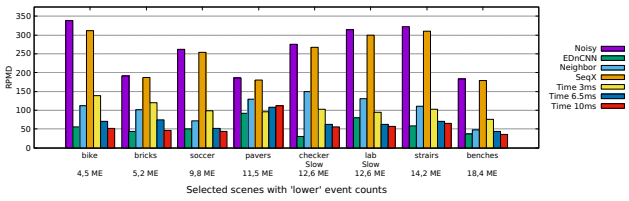


Figure 8: Comparison of EDnCNN benchmark results on `DVSNOISE20` dataset [4] (applying same calculation and averaging logic, smaller RPMD values indicate better denoising performance, ME $\hat{=}$ Million Events)

The denoising results obtained in this case by using the provided pre-trained EDnCNN network weights and the described spatiotemporal filter are given in Figure 8. For consistency with the results originally computed by Baldwin et al., the "relative plausibility measure of denoising" (RPMD) metric is also used here (for details see [4]).

The results are comparable to the denoising results on our dataset.

6. Conclusion

The central contribution of this paper is a neuromorphic vision dataset addressing the issues of DVS-based long time monitoring, especially in real outdoor scenarios. It consists of several hours of raw event data and a total of 47,878 regions of interest containing labels which were generated by the described processing chain. These labels address classical object classes as well as environmental interferences (like rain and shadows) which are included in DVS event streams in outdoor recordings, but have not been taken into account in available datasets so far.

With respect to the challenges that have to be considered in DVS-based processing pipelines we provide a quantitative comparison of denoising results utilizing different spatio-temporal filters on the provided dataset as well as on the DVSNOISE20 [4] dataset. In summary good denoising results can be achieved with respect to background events, while preserving a high amount of object events. However included events caused by environmental interferences continue to be a challenge. In further work, we will also address these challenges by means of higher level analysis for object detection and classification.

The dataset is made freely available to support and accelerate the development and deployment of fully DVS based processing pipelines in real world usage scenarios. Especially in these cases the advantages of Dynamic Vision Sensors like privacy aspects, low power consumption and a high dynamic range can be particularly beneficial.

Acknowledgment

This work was supported by the European Regional Development Fund under grant number EFRE-0801082 as part of the project "plsm" (`https://plsm-project.com/`).

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017. 4

[2] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner, and D. Modha. A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7388–7397, 2017. 2

[3] M. N. Asghar, N. Kanwal, B. Lee, M. Fleury, M. Herbst, and Y. Qiao. Visual surveillance within the eu general data protection regulation: A technology perspective. *IEEE Access*, 7:111709–111726, 2019. 3

[4] R. Wes Baldwin, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7, 8

[5] E. Barnoviciu, V. Ghenescu, S. Carata, M. Ghenescu, R. Mihaescu, and M. Chindea. Gdpr compliance in video surveillance and video processing application. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6, 2019. 3

[6] Y. Bi and Y. Andreopoulos. Pix2nvs: Parameterized conversion of pixel-domain video frames to neuromorphic vision streams. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1990–1994, 2017. 2

[7] S. Chen and M. Guo. Live demonstration: Celex-v: A 1m pixel multi-mode event-based sensor. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1682–1683, 2019. 3

[8] S. Chen, W. Tang, X. Zhang, and E. Culurciello. A 64 × 64 pixels uwb wireless temporal-difference digital image sensor. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 20(12):2232–2240, 2012. 3

[9] Wensheng* Cheng, Hao* Luo, Wen Yang, Lei Yu, Shoushun Chen, and Wei Li. Det: A high-resolution dvs dataset for lane extraction. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2019. 2

[10] Tobi Delbruck, Yuhuang Hu, and Zhe He. V2E: From video frames to realistic DVS event camera streams. *arxiv*, June 2020. 2

[11] G. Gallego, T. Delbruck, G. M. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 3, 4, 7

[12] G. P. Garca, P. Camilleri, Qian Liu, and S. Furber. pydvs: An extensible, real-time dynamic vision sensor emulator using off-the-shelf hardware. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2016. 2

[13] Marianne Gras. The legal regulation of cctv in europe. *Surveillance and Society*, 2, 01 2004. 3

[14] M. Guo, R. Ding, and S. Chen. Live demonstration: A dynamic vision sensor with direct logarithmic output and full-frame picture-on-demand. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 456–456, 2016. 3

[15] M. Guo, J. Huang, and S. Chen. Live demonstration: A 768 640 pixels 200meps dynamic vision sensor. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1, 2017. 3

[16] Shasha Guo, Lei Wang, Xiaofan Chen, Limeng Zhang, Ziyang Kang, and Weixia Xu. Seqxfilter: A memory-efficient denoising filter for dynamic vision sensors, 2020. 7

[17] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4

[18] Y. Hu, J. Binas, D. Neil, S. C. Liu, and T. Delbruck. Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, 2020. 2

[19] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in Neuroscience*, 10:405, 2016. 2

[20] J. Huang, M. Guo, and S. Chen. A dynamic vision sensor with direct logarithmic output and full-frame picture-on-demand. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, 2017. 3

[21] Cheng-Bin Jin, Trung Dung Do, Mingjie Liu, and Hakil Kim. Real-time action detection in video surveillance using a sub-action descriptor with multi-convolutional neural networks. *Journal of Institute of Control, Robotics and Systems*, 24:298–308, 03 2018. 1, 2

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 5

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 4

[24] CelePixel Technology Co. Ltd. Celex4 sdk. https://github.com/CelePixel/CeleX4-OpalKelly/, 2019. 3, 6

[25] Z. Luo, F. Branchaud-Charron, C. Lemaire, J. Konrad, S. Li, A. Mishra, A. Achkar, J. Eichel, and P. Jodoin. Mio-tcd: A new benchmark dataset for vehicle classification and localization. *IEEE Transactions on Image Processing*, 27(10):5129–5141, 2018. 1

[26] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois C Knoll. Neuromorphic benchmark datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in neurorobotics*, 13:38, 2019. 2

[27] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, 2011. 1, 2

[28] Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015. 2

[29] Vandana Padala, Arindam Basu, and Garrick Orchard. A noise filtering algorithm for event-based asynchronous change detection image sensors on truenorth and its implementation on truenorth. *Frontiers in Neuroscience*, 12:118, 2018. 7

[30] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16639–16652. Curran Associates, Inc., 2020. 2

[31] Mantini Pranav, Li Zhenggang, and Shah Shishir K. A day on campus - an anomaly detection dataset for events in a single camera. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 1, 2

[32] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 969–982. PMLR, 29–31 Oct 2018. 2

[33] Teresa Serrano-Gotarredona and Bernab Linares-Barranco. Poker-dvs and mnist-dvs. their history, how they were made, and other details. *Frontiers in Neuroscience*, 9:481, 2015. 2

[34] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[35] Corey Snyder and Minh Do. Streets: A novel camera network dataset for traffic flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10242–10253. Curran Associates, Inc., 2019. 1

[36] A. Vasudevan, P. Negri, B. Linares-Barranco, and T. Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 675–682, Los Alamitos, CA, USA, may 2020. IEEE Computer Society. 2