

N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation

Tobias Bolten¹^a, Christian Neumann¹^b, Regina Pohle-Fröhlich¹ and Klaus D. Tönnies²

¹Institute for Pattern Recognition, Hochschule Niederrhein, Krefeld, Germany

²Department of Simulation and Graphics, University of Magdeburg, Germany

{tobias.bolten, christian.neumann, regina.pohle}@hs-niederrhein.de, klaus@isg.cs.uni-magdeburg.de

Keywords: Dynamic Vision Sensor, Event Data, Instance Segmentation, Multi-Person Tracking, Dataset.

Abstract: Compared to well-studied frame-based imagers, event-based cameras form a new paradigm. They are biologically inspired optical sensors and differ in operation and output. While a conventional frame is dense and ordered, the output of an event camera is a sparse and unordered stream of output events. Therefore, to take full advantage of these sensors new datasets are needed for research and development. Despite their ongoing use, the selection and availability of event-based datasets is currently still limited. To address this limitation, we present a technical recording setup as well as a software processing pipeline for generating event-based recordings in the context of multi-person tracking. Our approach enables the automatic generation of highly accurate instance labels for each individual output event using color features in the scene. Additionally, we employed our method to release a dataset including one to four persons addressing the common challenges arising in multi-person tracking scenarios. This dataset contains nine different scenarios, with a total duration of over 85 minutes.

1 INTRODUCTION

Dynamic Vision Sensors (DVS) are optical sensors designed to replicate the basic neural architecture and operating principle of the human eye. Each pixel of a DVS detects and reacts completely asynchronously and independently to changes in brightness. In this process, each pixel generates and sends an output as soon as a brightness change above a set threshold value is detected. Therefore, unlike classic image sensors that operate with a fixed sampling rate, the output of a DVS is a completely data-driven stream of triggered output ‘events’. Each event contains information about

- the (x, y) position of the triggered pixel in the sensor array,
- a very high precision timestamp t of the time of occurrence, and
- an indicator for the direction of the detected brightness change.

The operation and output paradigm of the DVS results in technical advantages for tracking and segmentation tasks. Compared to classic systems, the

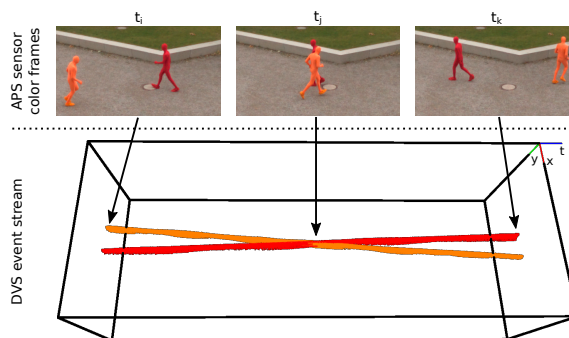




Figure 1: DVS event stream concept visualization. The high time resolution supports continuous tracking approaches by using high-quality segmentations.

DVS signal contains significantly less redundant information, as only changes are recorded. For segmentation tasks it is advantageous that static components do not need to be processed. Due to the high time resolution of the DVS, the resulting 3D (x, y, t) space-time event cloud provides an almost continuous signal for moving objects (see Figure 1). A high-quality segmentation thus supports the future development of DVS-based tracking methods. However, this requires publicly available datasets that include these segmentation annotations as well as object tracking scenarios and challenges.

^a <https://orcid.org/0000-0001-5504-8472>

^b <https://orcid.org/0000-0002-0871-8629>

In the context of multi-object tracking (MOT), a dataset should include the following common challenges (Islam et al., 2015; Xu et al., 2019; Luo et al., 2021):

1. object occlusions (through infrastructure as well as by other persons in the scene)
2. similar appearance and body shape of recorded persons
3. included changes in pose and movement patterns (e.g. kneeling, standing, walking, and running)
4. interactions among multiple persons (including abrupt changes in movement direction and speed)
5. included objects in different sizes

In addition, to take advantage of the high temporal resolution of the sensor, this dataset must provide the instance annotations for each event of the DVS output stream. Due to the novelty of the sensor itself, the availability of pure event-based datasets is limited compared to the frame-based domain. To the best of the authors' knowledge, currently no DVS dataset fulfills these requirements.

In this paper, we present a technical setup for recording and also publish a dataset which aims to provide instance annotations on DVS data. In summary, we contribute:

- a detailed description of a hardware setup and corresponding software processing pipeline allowing the acquisition of multi-person DVS recordings with high-quality per-event instance labels
- a ready-to-use and publicly available multi-person DVS dataset that includes the aforementioned requirements within an outdoor recording setup.

Section 1.1 summarizes related work in terms of existing datasets and event-based approaches. The requirements for our recording setup are specified in Section 2. Subsequently, the hardware setup used is described in Section 3. The proposed software pipeline for generating object annotations is presented in detail in Section 4. Section 5 outlines the details and statistics of the provided dataset.

1.1 Related Work

Early event-based tracking approaches were developed and evaluated using basic datasets containing, for example, objects of simple geometric shape moving (c.f. the *shape* scenes from (Mueggler et al., 2017b)). These types of scenes were used to detect and track features such as corners (Mueggler et al., 2017a; Alzugaray and Chli, 2018). In the context of this paper, however, more complex event-based

datasets concerning person detection and tracking are of greater interest.

In (Jiang et al., 2019) an approach for person detection in the application area of traffic surveillance was presented. For this purpose, a custom dataset was used which only has a total length of ≈ 14 seconds. In a very similar application context, a person detection was also implemented in (Chen et al., 2019). This was followed by the publication of a dataset in (Miao et al., 2019). The detection part of this dataset consists of only twelve short-length sequences. In (Ojeda et al., 2020) and (Bisulco et al., 2020) two approaches for filtering the event stream are presented with the goal of implementing person detection close to hardware. For this, a DVS dataset consisting of several hundred sequences is used. Yet, this dataset is not publicly available.

The task of multi-person tracking is directly addressed in (Piątkowska et al., 2012). Due to the lack of available ground-truth annotations, their approach was only tested on a very limited set of data which were manually labeled. With the work of (Hu et al., 2016), DVS benchmark datasets have been published. This includes a part that explicitly relates to object tracking. In this case, though, only the tracking of a single object is considered. In (Camuñas-Mesa et al., 2018) challenges arising from object occlusion are addressed for an event-based tracking. The real-world capabilities of their approach was tested on a multi-person tracking scenario. But the qualitative results were only considered on the basis of a short scene.

In summary, it can be stated that there are currently only a few event-based datasets available within the context of person detection and tracking. Furthermore, label annotations included in those datasets are not sufficiently discriminative. In large-scale object detection datasets such as the *GENI* automotive dataset (de Tournemire et al., 2020), object labels are often specified only in the form of bounding boxes. With (Alonso and Murillo, 2019; Bolten et al., 2021), there are event-based datasets that provide annotations for semantic segmentation. Nevertheless, to fulfill the initial requirements in the context of multi-person tracking, an instance-level annotation is required. Currently, no suitable event datasets exist that satisfy this.

In conventional frame-based computer vision, datasets exist that provide annotations beyond bounding boxes for tracking and segmentation (Voigtlaender et al., 2019). With work like (Hu et al., 2021), approaches do exist to convert frame-based datasets into the event-based domain. However, this conversion does not reflect the non-ideal sensor characteristics fully. Therefore, we propose a procedure that

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>

allows the creation of a native DVS dataset with an automated generation of labels requiring only minimal manual work.

Like (Marcireau et al., 2018), our approach is also based on exploiting color features within the scene. Compared to their hardware setup, which consists of three DVS and an optical beam splitter, our setup is simpler and does not impose constraints on the recorded spectrum. A detailed description of our used setup is given in Section 3.

2 LABEL EXTRACTION

A meaningful dataset requires accurate labels. In the case of instance labeling of event data, manual annotation is not efficient. We solved this problem by introducing additional information in a way so that the original data is not influenced. More specifically, a DVS is usually based on CMOS technology. A CMOS image sensor typically is most sensitive to near-infrared light. DVS with color filter matrices exist only as prototypes and do not feature the higher image resolution of newer DVS models (Moeys et al., 2018; Taverni et al., 2018). Thus, for practical purposes DVS are not capable of recording color information. We are using this circumstance to encode the information which event belongs to a given person in the color of the clothes of the persons itself.

In order to record the color information, a second frame-based color camera, referred to as an active pixel sensor (APS) camera, is required. Assuming proper recordings, label data with high accuracy can be extracted from the color frames.

2.1 Color Features

The color features are generated by single-colored full-body skin-tight garments. The following specifications are important for this:

1. A single color per suit is required because person instances will be separated by color hue.
2. The suit must cover the whole body so that color information is available for all event data triggered by a person.
3. The suit should be skin-tight so that the silhouette of a person is not larger than normal.

The color of the suits changes the recorded intensities of the DVS only slightly. Synthetic fabrics tend to be strongly NIR-reflective. Independent of the visible color, the reflected NIR-light dominates the spectral response of the garment. It is thus able to trigger DVS events for any chosen color.

Initial experiments showed that fabrics can be distinguished by color hue even when the separation in hue is small (see Figure 5 and the description in Section 4.2). This enables the use of many colors and therefore many person instances can be distinguished simultaneously.

However, there are multiples sources of color artifacts that must be considered:

1. The lens introduces chromatic aberrations, i.e. phantom colors towards the edges of the image area, that can only be partially corrected.
2. The camera records color information through a color filter matrix. Only three colors, red, green, and blue, are recorded. The real source color is a mixture of these and colors first needs to be reconstructed during demosaicing. During this process errors are introduced near borders because neighboring color signals could be combined while the sources were separated in reality. This also introduces new colors that are not part of the real scene. As separation of persons by color hue is the basis of our work, erroneous color information hinders our method.
3. A third source of errors is the on-chip *binning* process of the camera sensor. The effect that *binning* has on resulting colors is similar to the previous point. By aggregating 2×2 pixels it is possible to get phantom colors not included in the real scene.

Also over-exposure in each color channel must be avoided. Single-colored suits can quickly lead to over-exposure in a single channel. It is recommended to deliberately under-expose the scene so that no required information is lost.

2.2 Environmental Influences

Data acquisition in an outdoor environment is naturally influenced by the environment itself. One important aspect is the lighting of the scene. Direct sunlight is not preferred because the color of the direct sunlight itself is different from the color of diffusely reflected light. In consequence, the parts in direct sunlight will need a different white balance than parts in the shadows. The information about which parts are in direct sunlight is not known precisely and thus the necessary distinction can not be made. Overcast is preferred for its diffuse scene lightning effect.

Another practical problem are airborne particles. Acquisition during precipitation is not useful because the sensitivity in detecting changes in brightness and the temporal resolution of a DVS is high enough to image the rain drops themselves. Wet scenes should

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>

be avoided due to the possibility of unwanted reflections and apparent color shift of wet surfaces of some materials. Thus, dry weather without larger airborne particles is preferred. This includes also insects and pollen, as these environmental influences would also be included in the DVS signal.

3 HARDWARE SETUP

In order to enable accurate processing of event and color information, several aspects and requirements regarding the hardware setup need to be considered. This includes the selection of the sensors used as well as their mounting in a stereoscopic setup to optimize the perspective mapping. Furthermore, the colored garments used to incorporate the color features must also be taken into account.

3.1 Sensors

The DVS is a CeleX-IV built by CelePixel (Guo et al., 2017). The sensor is combined with an 8mm wide-angle lens from Computar¹. The DVS is connected via USB 3 to a notebook that stores the event data on an external solid state drive. Previous experiments showed that event noise is strongly dependent on sensor temperature (Nozaki and Delbruck, 2017; Berthelon et al., 2018). We applied a cooling system based on Peltier elements in order to stabilize the operating temperature at a low level.

Our APS camera is based on a Sony IMX477 CMOS image sensor. For the color frames, it is necessary to gather the data in a lossless manner. Our requirements are minimal frame drops, maximal resolution, full color information, and maximal frames per second (fps). State-of-the-art video compression like H.264 allocates few bits to color information. The resulting hue values are not usable for our purposes. It is possible to record a raw video stream from the IMX477 sensor. The video stream features a resolution of 4056×3040 px with 12-bit per pixel. The resulting data rates are a challenge for realtime applications. Multiple limitations arise here:

- The camera interface on the APS sensor board is not capable of transmitting the full resolution in uncompressed form at the desired rate of frames per second. Thus, *binning* was activated which in turn halves the image dimensions to 2028×1520 px.

¹Computar MEGAPIXEL V0814-MP, f=8mm, f/1.4-f/16, 1 inch, C-Mount

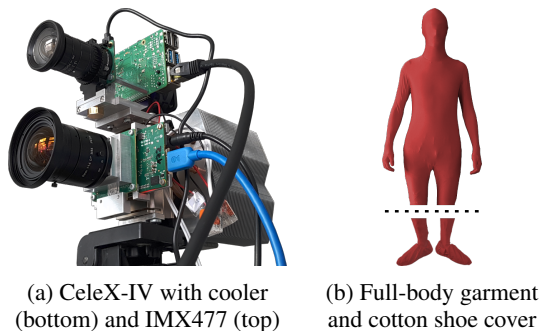


Figure 2: Sensor and color garment setup.

- The accumulated data rates of the APS and DVS as well as the storage on an external USB media exceed the capacity of a single computer’s USB bus. Therefore a dedicated Raspberry Pi 4 was used to acquire the APS frames and store them to another external solid state drive.

The field of view was matched as close as possible by the choice of the lens. The IMX477 was combined with a 4mm wide-angle lens from Edmund Optics²

Both sensors together form a stereoscopic camera. The cameras are stacked vertically as parallax is less apparent along the vertical axis with cameras looking at the scene from an elevated position. Also, the lenses were mounted as close to each other as possible. The physical mount was milled from aluminum alloy and placed on a solid surface so that the likelihood of tripod shake triggering unwanted events is minimized. The optical axis were adjusted towards the same point in the center of the scene so that lateral overlap of the resulting views is maximal. The resulting hardware setup is shown in Figure 2a.

3.2 Color Garments

The full-body garment is made up of two parts. The complete body, including head, hands, and feet, is covered by a Morphsuit – a commercially available Spandex-based suit. In order to guarantee a natural walking pattern, all persons were allowed to wear everyday shoes. The shoes are covered using separate covers sewed from colored cotton fabric. An example of a complete garment is depicted in Figure 2b.

4 SOFTWARE PIPELINE

The software processing pipeline for generating the label masks as well as their propagation to the DVS

²Edmund Optics TECHSPEC® UC Series #33-300, f=4mm, f/1.8-f/11, 1/2 inch, C-Mount.

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>

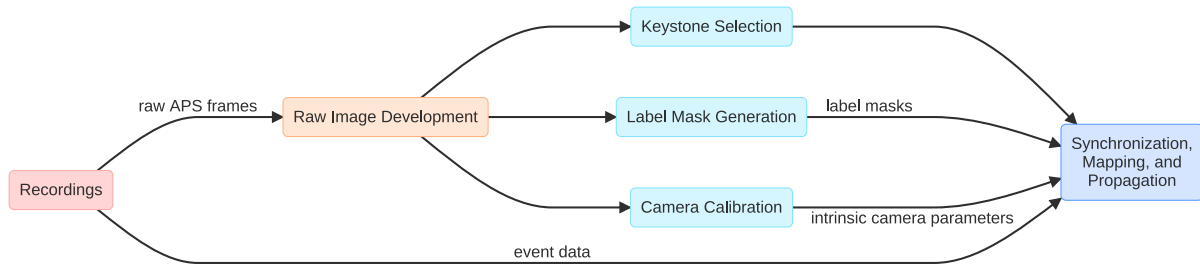


Figure 3: Overview of software pipeline used for dataset generation.

data consists of several stages which are explained in the following. An overview of the process is given in Figure 3.

4.1 Raw Image Development

To minimize color artifacts and inaccuracies, especially in hue, the frame-based APS recordings from the IMX477 sensor are performed as lossless as possible. This means that a full digital photo development must be applied to develop appropriate color images from the raw data.

In order to minimize the required data bandwidth of the sensor, the acquisition is performed in a packed data format. For every two 12-bit pixels, three bytes are combined by the sensor. After unpacking the individual pixels from this structure, a black level correction is performed and a color image is generated by using an edge-aware demosaicing algorithm on the Bayer-filtered data. Following clipping and normalization of the data, a color correction matrix is applied to transfer the sensor-specific color values into a standardized, device-independent color space. Finally a gamma correction is applied and the resulting frame is cropped to the resolution of 2000×1500 px to remove information-free stride pixels and color artifacts at the borders from the image.

In a second step, chromatic aberrations are corrected. For this, the image is further adapted with an open-source photo editor³.

4.2 Label Mask Generation

Label masks are automatically generated based on the developed APS frames by means of color hue segmentation.

First, regions of interest are marked by hand (see Figure 4). Areas of different illumination were included. This is a countermeasure against the influences described in Section 2.2. After conversion into HSV color space, we get multiple color clusters per



Figure 4: Example of one-time manually annotated color areas for the performed clustering.

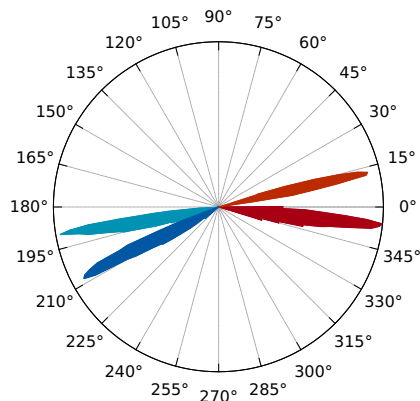


Figure 5: Polar histogram of garment colors as contained in annotations on developed APS frames used for dataset.

garment color. The clusters can be visualized in a polar histogram with 360 bins (see Figure 5). These clusters are further processed to give a single centroid in hue for each garment color.

For this purpose, an agglomerative clustering is computed on the hue values. Clusters with differences in hue below a given threshold are merged. This threshold is increased as long as the number of resulting clusters is lower than the number of garment colors used during recording. It is assumed that a normal distribution is suited to model potential color variations, e.g. due to lighting. A normal distribution is fit to each cluster. The calculated mean hue of each cluster is used for the hue segmentation in the next step. The hue range in HSV color space is circular. Care must be taken to use circular mean for all computations involving hue.

³<https://www.darktable.org/>

The hue segmentation is computed by calculating the differences to the centroids in hue of each cluster and then applying a threshold α . At this point, we have selected one circular sector for each color in HSV color space. We can further narrow the decision regions by incorporating saturation and value. The idea is to exclude all colors that cannot origin from the color garment in the given setup. We remove all pixels with a saturation $\leq \sigma$ and a value $\leq \varphi$. The thresholds must be selected so that as little as possible from the regions of interest is cut away. At last a morphological opening closes small holes. Now we have separate binary mask for each garment color.

Additional care was taken to prevent multiple labels per pixel and removal of false positives. When the resulting label mask is set for multiple colors at one position only one color is kept. Centroids' hue values are used in ascending order to define priorities. Most false positives are removed by applying a connected component analysis and suppressing all connected components with a size of less than τ .

4.3 Synchronization, Mapping & Propagation

Several steps are necessary to project the obtained label masks onto the DVS event stream. A priori, intrinsic camera parameters must be estimated for both sensors of the setup. The required points can be extracted using a chessboard moving slowly and applying corner detection to it.

It is important to temporally synchronize the views so that label masks derived from the APS camera correspond to the view of the DVS. The recording of raw data from the APS color sensor is technically limited to 40fps by the hardware setup used⁴. For further processing, the continuous data stream of the DVS is therefore also divided into sections of the length of 25ms. The resulting APS frames and the time windows of the DVS are then synchronized to each other on the basis of system clock timestamps corresponding to their acquisition time. The clocks of the systems used are adjusted via NTP. This limits the error over time to a small value. Additionally, an initial synchronization at the start of every recording is manually performed using a visual cue. This visual cue is used to compensate most of the absolute error between the system timestamps.

For mapping, the generated label mask is first undistorted with the determined camera parameters of

⁴Using the 12-bit packed data format at a sampling rate of 40fps results in an approximate data rate of $178 \frac{\text{MB}}{\text{s}}$ which must be continuously transferred and stored.

the APS. This process step is illustrated exemplarily in Figure 6a and 6b. Then the mask is projected onto the DVS, changing the field of view and image resolution (see Figure 6c). For this purpose, a homography is determined on the basis of manually selected points on the ground plane of the scene. This is then used to warp the mask in perspective.

The redistorted instance labels (see Figure 6d) are finally propagated to the plain DVS event stream. All events within a synchronized time window are assigned with the label that the mask has at each corresponding spatial position. The final result is thus a per-event instance label annotation on the DVS output stream (see Figure 6e).

4.4 Limitations

There is one notable limitation. The label mask generation based on the APS frames works for both, moving and standing persons. In contrast, a DVS ideally only generates event data for regions with movement. In consequence, noise of the DVS is labeled as belonging to a person when a person is not moving.

The label is wrong when looking at event data itself because the events were not triggered by the persons themselves. But when viewed from the standpoint of tracking algorithm the label is correct, because a person cannot disappear in a scene. Therefore, we decided not to further post-process the label masks to address this issue.

5 DATASET: N-MuPeTS⁵

First, we give an overview of what is included in the dataset and briefly explain every annotation used. After that we discuss multiple statistics derived from the dataset.

During recording in an unconstrained outside environment, interference can not be completely avoided. Every deviation from a perfect sequence of actions must be edited during post-processing. We decided to mark those sequences which are of lower quality and split the recording at the point in time when a change in quality occurs. This process was done manually to ensure high accuracy. We observed a number of issues, including both, recording artifacts and post-processing errors. We distinguished three classes of quality by the impact of the issues on usability for tracking applications. From worst to best, the quality classes are:

⁵Neuromorphic-Multi-Person Tracking and Segmentation Dataset

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>

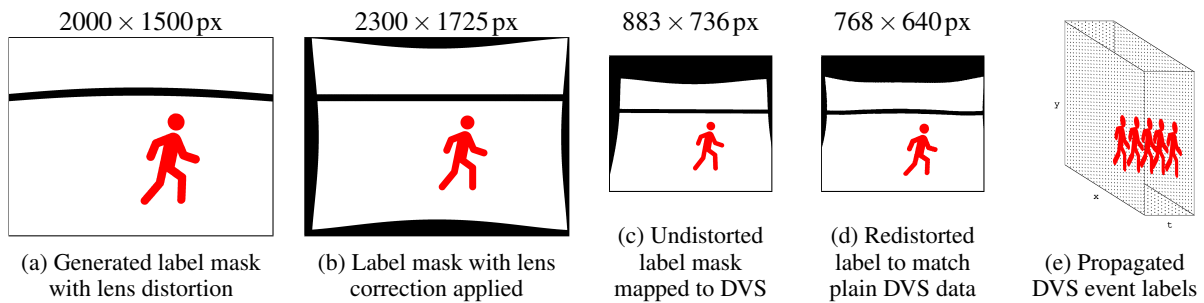


Figure 6: Processing steps to map and propagate label masks to DVS event stream.

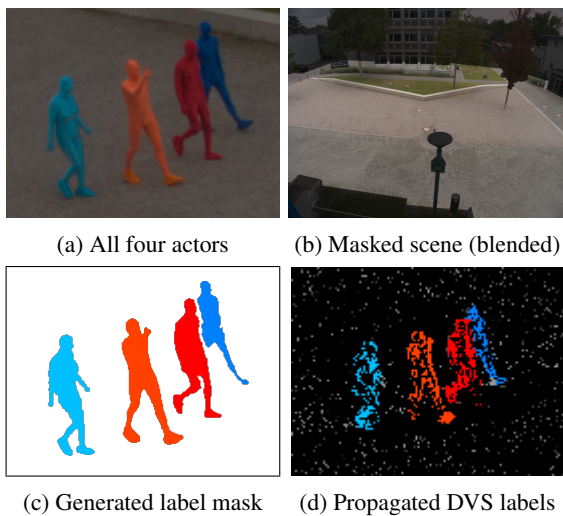


Figure 7: Setup during dataset recording (top row) and corresponding labels (bottom row).

Quality 3 corresponds to these major issues:

- uninvolved persons in scene
- cars moving in scene
- wildlife in scene, i.e. birds

Quality 2 corresponds to these minor issues:

- unwanted occlusion, i.e. tree trunks
- one or more person outside of static mask
- incomplete masks, i.e. intersection with static mask (see Section 5.3)

Quality 1 includes all remaining sequences, i.e. sequences without any of the aforementioned issues.

Quality 2 and 3 can be used to get longer sequences (c.f. supplement to this paper). In the following, we present quality 1 exclusively. The complete dataset, including sequences of all three qualities, is available publicly and free of charge⁶.

⁶<http://dnt.kr.hsnr.de/DVS-NMuPeTS/>

The supplementary material to this paper is also available for download there.

Table 1: Scenarios.

pattern	primary parameter
background (empty scene)	–
single person	speed
crossing paths	angles of paths, speed
parallel paths (same directions)	speed, distance
parallel paths (opposing directions)	speed, distance
occlusions	speed, pose
meeting & parting	direction
random path	speed
helical path	–

5.1 Scenarios

We defined a set of scenarios as a protocol for recording and guidance to our actors. Where possible, we repeated the scenario for each person and all possible numbers of persons. The scenarios are listed in Table 1. The order is mostly chronological and corresponding to sequence indices. For the majority of the scenarios, we introduced annotations with a corresponding name. The annotations are not exclusive for certain sequences, e.g. CROSSING will occur frequently during ‘crossing paths’ but also anytime else. Thus, the scenarios increase the frequency of their corresponding annotations for their duration.

For the generation of label masks we used $\alpha = 10^\circ$, $\sigma = 50\%$, $\varphi = 10\%$, and $\tau = 10\text{px}$.

5.2 Persons & Garment Colors

The recorded persons are pseudonymized in the provided data by using the color of the suit they were wearing (see Table 2 as well as Figure 7a and 8). Some colors are less practical than others in a given environment. In our case, vegetation tends to be green-yellowish. Red and blue are well separated.

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>

Table 2: Technical specifications of participating persons.

garment color	size	sex	weight
red	1.84 m	m	70 kg
orange	1.82 m	m	82 kg
cyan	1.74 m	f	80 kg
blue	1.80 m	m	65 kg



Figure 8: Overview of the fabrics used. First row: cotton shoe covers, second row: Morphsuits. Sorted by hue, from left to right, i.e. red, orange, cyan, and blue.

The choice of orange and cyan, which are close in hue (see Figure 5), is due to unavailability of differently colored suits during the time of dataset creation. Figure 7c shows an automatically generated label mask in the used setup, while Figure 7d shows the result of corresponding mapping and propagation to the DVS view. The supplement provides further examples of generated label masks for a qualitative overview.

5.3 Static Mask

We used a binary mask to prevent vegetation as well as uninvolved persons and objects, as described under ‘Quality 3’, from triggering false positives during label mask generation. Some sequences needed to be split and marked as quality 2 instead of quality 1 because one or more person is masked out by the static mask.

This static mask is only applied to the color frames from the APS camera. In the dataset, all event data is unmasked. You can see the mask applied in Figure 7b.

5.4 Annotations

Annotations were made manually and are available per-color.

The scenario description, used to brief the actors, and the annotations, describing all included activities, are directly related to the main challenges of MOT defined in the introduction. We will now briefly discuss the mapping and explain our annotations.

5.4.1 Background

One annotation not resembling a problem in itself is **BACKGROUND**, i.e. an empty scene. This provides

separate recordings only including sensor noise. In quality 2, some foreign activity can occur.

BACKGROUND no person with colored garment is in the scene.

5.4.2 Object Occlusion

Object occlusion can be further divided into occlusion with infrastructure and occlusion between persons. Occlusion between persons can happen whenever two or more persons are in the scene. It is very frequent during **CROSSING** and **SIDEBYSIDE** (see Section 5.4.5).

OCCLUSION one or more persons are behind an obstacle.

5.4.3 Similar Shape

All actors are of comparable size and weight. The person representing **CYAN** naturally has a different silhouette while the remaining persons have a very similar figure. This makes distinction by size impracticable and provides a greater challenge for tracking algorithms.

RED, ORANGE, CYAN, BLUE person with specified garment color is in scene.

5.4.4 Pose & Movement

Most of the time the actors were upright and walking. For short durations they were in differing poses. **EXERCISING** is a collection of movement patterns found in sports. This includes push-ups and burpees (including jumping). These represent movements rarely observed in pedestrian monitoring scenes with high rates of change in movement speed and thus may pose a challenge for tracking algorithms.

STANDING, WALKING, and RUNNING correspond to three intervals of speed during bipedalism.

RANDOM is a special case where one actor tries to move as unpredictable as possible while including changes of speed and course. This aims at stressing physical motion models included in many tracking algorithms.

EXERCISING one or more persons are doing exercise like activity.

KNEELING, STOOPED, WAVING one or more persons are in the specified pose.

STANDING, WALKING, RUNNING one or more persons are moving the specified manner.

RANDOM one or more persons are abruptly changing direction and speed, including backwards movement.

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>

5.4.5 Interaction

Crossing paths require two persons, but the number of possible variations is limited. We consider the following patterns to be relevant:

1. person A and person B walk past each other (CROSSING)
2. person A and person B walk in the same direction (SIDEBYSIDE)
3. person A or person B change their direction at the point of intersection (MEET)

Parameters are the angle towards the camera and between paths, and the speed. With three persons the possible number of variations increases drastically. For this dataset, we managed to gather four persons.

CROSSING the paths of two or more persons are crossing in temporal proximity.

MEET two or more persons are meeting and parting again, paths are converging then diverging.

SIDEBYSIDE two or more persons are moving side by side at the same speed.

HELIX two persons try to circle around each other so that in a (x, y, t) -diagram, i.e. a 3D plot of the resulting event point cloud, their paths resemble a helical path.

5.4.6 Different Sizes

As mentioned before, the actors are of similar size. Still the projected size varies greatly with distance to the camera due to perspective. While a person in the center of the scene has an projected size of approximately 52px in the DVS event stream, at the far end of the scene a person is only 29px in size. Sequences with persons at the far end of the scene are marked with FAR.

In conclusion, persons are included in the dataset in multiple sizes. Of course, the variation in size corresponds directly to the position along the vertical axis in the stream.

FAR one or more persons are near the distant footway in the scene.

5.5 Statistics

The dataset subset constituting quality class 1 is summarized in Tables 3, 4, and 5. In these tables, all durations are given in seconds and rounded to the nearest integer. The supplement contains detailed information for all quality classes and annotations. Overall, the cumulative duration of sequences in quality 1 is

Table 3: Cumulative durations per color combination in quality class 1.

RED	ORANGE	CYAN	BLUE	cumulative duration	
				per color combination	sum per person count
					300
•				313	
	•			158	
		•		215	882
			•	195	
•	•			213	
•		•		41	
•			•	125	701
	•	•		236	
	•		•	76	
		•	•	9	
•	•	•		496	
•	•		•	123	687
•		•	•	51	
	•	•	•	17	
•	•	•	•	350	350

Table 4: Duration statistics per annotation in quality class 1

annotation	number of sequences	mean duration	cumulative duration
RED	137	12.5	1713
ORANGE	123	13.6	1671
CYAN	101	14.0	1416
BLUE	77	12.3	946
BACKGROUND	46	6.5	300
STANDING	96	3.3	312
WALKING	441	5.1	2259
RUNNING	107	4.7	504
RANDOM	18	8.9	160
HELIX	9	7.2	64
FAR	35	4.5	157

approximately 56% of the total dataset which equates to a length of ≈ 2920 seconds.

Table 3 summarizes the durations recorded per color combination. It can be noted that all possible color combinations are included. The last column contains the cumulated durations for 0, 1, 2, 3, and 4 actors active at the same time. Recordings with up to two or three persons each contribute 24% of the total dataset duration. Scenes with one person ($\approx 30\%$)

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>

Table 5: Occurrence statistics per annotation in quality class 1.

annotation	number of occurrences
OCCLUSION	136
EXERCISING	9
KNEELING	13
STOOPED	18
WAVING	9
CROSSING	212
MEET	48
SIDE BY SIDE	94

and with all four persons ($\approx 12\%$), along with empty background scenes, form the remaining components.

Table 4 gives an overview of the amount of sequences and their duration in relation to the assigned annotations. The cumulative duration is included in a separate column for convenience. Due to the aforementioned rounding, small discrepancies can occur. Annotations describing short events in time are instead counted, see Table 5.

Considering the cumulative duration each actor is present, RED, ORANGE and CYAN are included in equal shares. Actor BLUE is slightly less frequent.

According to the natural movement pattern of humans, the annotation WALKING is included significantly more often than others.

6 CONCLUSION

Currently, there is still a lack of publicly available event-based datasets. This is an obstacle in the development and evaluation of event-based image processing applications. We have described a software-processing pipeline and its associated hardware setup for an automated derivation of multi-person instance annotations within DVS recordings. By exploiting color features, highly accurate annotations are generated. This is even the case in error-prone scenarios including intersecting movements and occlusions.

Additionally, we published an already ready-to-use dataset that includes the scenarios and challenges that arise in the context of multi-person tracking applications. This dataset can be used for further development of event-based algorithms. As this dataset provides annotations on the level of the individual output events of the DVS, there are no event encoding constraints which must be considered. This includes approaches that operate directly on the 3D (x, y, t) space-time event cloud of the DVS.

Potential challenges supported by this dataset include therefore, for example, the application and evaluation of 3D cloud-based algorithms for instance segmentation (Yang et al., 2019; Zhao and Tao, 2020) into the event-based vision. The transition of graph-based approaches, such as from object detection (Mondal et al., 2021) and recognition (Li et al., 2021) within DVS data, to tracking is also a supported and an interesting topic for further work.

ACKNOWLEDGEMENTS

We thank Hans-Günter Hirsch for his support and participation in the recording of the dataset.

Funding

This work was supported by the Federal Ministry for Digital and Transport under the mFUND program grant number 19F1115A as part of the project ‘TUNUKI’.

REFERENCES

- Alonso, I. and Murillo, A. C. (2019). EV-SegNet: Semantic Segmentation for Event-Based Cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1624–1633.
- Alzugaray, I. and Chli, M. (2018). Asynchronous Corner Detection and Tracking for Event Cameras in Real Time. *IEEE Robotics and Automation Letters*, 3(4):3177–3184.
- Berthelon, X., Chenegros, G., Finateu, T., Ieng, S.-H., and Benosman, R. (2018). Effects of Cooling on the SNR and Contrast Detection of a Low-Light Event-Based Camera. *IEEE Transactions on Biomedical Circuits and Systems*, 12(6):1467–1474.
- Bisulco, A., Cladera Ojeda, F., Isler, V., and Lee, D. D. (2020). Near-Chip Dynamic Vision Filtering for Low-Bandwidth Pedestrian Detection. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 234–239.
- Bolten, T., Pohle-Fröhlich, R., and Tönnies, K. D. (2021). DVS-OUTLAB: A Neuromorphic Event-Based Long Time Monitoring Dataset for Real-World Outdoor Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1348–1357.
- Camuñas-Mesa, L. A., Serrano-Gotarredona, T., Ieng, S.-H., Benosman, R., and Linares-Barranco, B. (2018). Event-Driven Stereo Visual Tracking Algorithm to Solve Object Occlusion. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):4223–4237.

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

- Chen, G., Cao, H., Ye, C., Zhang, Z., Liu, X., Mo, X., Qu, Z., Conradt, J., Röhrbein, F., and Knoll, A. (2019). Multi-Cue Event Information Fusion for Pedestrian Detection With Neuromorphic Vision Sensors. *Frontiers in Neurobotics*, 13:10.
- de Tournemire, P., Nitti, D., Perot, E., Migliore, D., and Sironi, A. (2020). A Large Scale Event-based Detection Dataset for Automotive. *arXiv*, abs/2001.08499.
- Guo, M., Huang, J., and Chen, S. (2017). Live demonstration: A 768×640 pixels 200Meps dynamic vision sensor. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1.
- Hu, Y., Liu, H., Pfeiffer, M., and Delbruck, T. (2016). DVS Benchmark Datasets for Object Tracking, Action Recognition, and Object Recognition. *Frontiers in Neuroscience*, 10.
- Hu, Y., Liu, S.-C., and Delbruck, T. (2021). v2e: From Video Frames to Realistic DVS Events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1312–1321.
- Islam, M. Z., Islam, M., and Rana, M. S. (2015). Problem Analysis of Multiple Object Tracking System: A Critical Review. *International Journal of Advanced Research in Computer and Communication Engineering*, 4:374–377.
- Jiang, Z., Xia, P., Huang, K., Stechele, W., Chen, G., Bing, Z., and Knoll, A. (2019). Mixed Frame-/Event-Driven Fast Pedestrian Detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8332–8338.
- Li, Y., Zhou, H., Yang, B., Zhang, Y., Cui, Z., Bao, H., and Zhang, G. (2021). Graph-based Asynchronous Event Processing for Rapid Object Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 914–923.
- Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., and Kim, T.-K. (2021). Multiple Object Tracking: A Literature Review. *Artificial Intelligence*, 293:103448.
- Marcireau, A., Jeng, S.-H., Simon-Chane, C., and Benosman, R. B. (2018). Event-Based Color Segmentation With a High Dynamic Range Sensor. *Frontiers in Neuroscience*, 12.
- Miao, S., Chen, G., Ning, X., Zi, Y., Ren, K., Bing, Z., and Knoll, A. (2019). Neuromorphic Vision Datasets for Pedestrian Detection, Action Recognition, and Fall Detection. *Frontiers in Neurobotics*, 13:38.
- Moeyes, D. P., Corradi, F., Li, C., Bamford, S. A., Longinotti, L., Voigt, F. F., Berry, S., Taverni, G., Helmchen, F., and Delbruck, T. (2018). A Sensitive Dynamic and Active Pixel Vision Sensor for Color or Neural Imaging Applications. *IEEE Transactions on Biomedical Circuits and Systems*, 12(1):123–136.
- Mondal, A., R, S., Giraldo, J. H., Bouwmans, T., and Chowdhury, A. S. (2021). Moving Object Detection for Event-based Vision using Graph Spectral Clustering. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 876–884.
- Mueggler, E., Bartolozzi, C., and Scaramuzza, D. (2017a). Fast Event-based Corner Detection. In Tae-Kyun Kim, Stefanos Zafeiriou, G. B. and Mikolajczyk, K., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 33.1–33.11. BMVA Press.
- Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., and Scaramuzza, D. (2017b). The Event-Camera Dataset and Simulator: Event-based Data for Pose Estimation, Visual Odometry, and SLAM. *The International Journal of Robotics Research*, 36(2):142–149.
- Nozaki, Y. and Delbruck, T. (2017). Temperature and Parasitic Photocurrent Effects in Dynamic Vision Sensors. *IEEE Transactions on Electron Devices*, 64(8):3239–3245.
- Ojeda, F. C., Bisulco, A., Kepple, D., Isler, V., and Lee, D. D. (2020). On-Device Event Filtering with Binary Neural Networks for Pedestrian Detection Using Neuromorphic Vision Sensors. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3084–3088.
- Piątkowska, E., Belbachir, A. N., Schraml, S., and Gelautz, M. (2012). Spatiotemporal multiple persons tracking using dynamic vision sensor. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 35–40.
- Taverni, G., Paul Moeyes, D., Li, C., Cavaco, C., Motsnyi, V., San Segundo Bello, D., and Delbruck, T. (2018). Front and Back Illuminated Dynamic and Active Pixel Vision Sensors Comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681.
- Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., and Leibe, B. (2019). MOTs: Multi-Object Tracking and Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7934–7943.
- Xu, Y., Zhou, X., Chen, S., and Li, F. (2019). Deep Learning for Multiple Object Tracking: A Survey. *IET Computer Vision*, 13(4):355–368.
- Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., and Trigoni, N. (2019). Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Zhao, L. and Tao, W. (2020). JSNet: Joint Instance and Semantic Segmentation of 3D Point Clouds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12951–12958.

This is a self-archived version of the paper: Bolten, T.; Neumann, C.; Pohle-Fröhlich, R. and Tönnies, K. (2023). **N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation**. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-634-7, ISSN 2184-4321, pages 290-300

The final version is available online at: <http://dx.doi.org/10.5220/0011680300003417>