NeuroAED: Towards Efficient Abnormal Event Detection in Visual Surveillance With Neuromorphic Vision Sensor

Guang Chen, *Member, IEEE*, Peigen Liu, Zhengfa Liu, Huajin Tang, *Senior Member, IEEE*, Lin Hong, Jinhu Dong, Jörg Conradt, *Senior Member, IEEE*, and Alois Knoll, *Senior Member, IEEE*

***Abstract*— Abnormal event detection is an important task in research and industrial applications, which has received considerable attention in recent years. Existing methods usually rely on standard frame-based cameras to record the data and process them with computer vision technologies. In contrast, this paper presents a novel neuromorphic vision based abnormal event detection system. Compared to the frame-based camera, neuromorphic vision sensors, such as Dynamic Vision Sensor (DVS), do not acquire full images at a fixed frame rate but rather have independent pixels that output intensity changes (called *events*) asynchronously at the time they occur. Thus, it avoids the design of the encryption scheme. Since *events* are triggered by moving edges on the scene, DVS is a natural motion detector for the abnormal objects and automatically filters out any temporally-redundant information. Based on this unique output, we first propose a highly efficient method based on the event density to select activated *event* cuboids and locate the foreground. We design a novel *event*-based multiscale spatio-temporal descriptor to extract features from the activated *event* cuboids for the abnormal event detection. Additionally, we build the NeuroAED dataset, the first public dataset dedicated to abnormal event detection with neuromorphic vision sensor. The NeuroAED dataset consists of four sub-datasets: *Walking, Campus, Square, and Stair* dataset. Experiments are conducted based on these datasets and demonstrate the high efficiency and accuracy of our method.**

***Index Terms*— Abnormal event detection, video surveillance, optical flow, event based descriptors, neuromorphic vision sensor.**

I. Introduction

WITH the increasing awareness of the public security, abnormal event detection (AED) in video surveillance plays a more and more important role in ensuring public safety. A typical method to detect anomaly event is to detect patterns in video scenes that do not agree with the established normality (see Fig. 1). With the rapid development of computer vision technologies, AED has a wide range of applications, such as crowd surveillance [1], public security [2], traffic monitoring [3], and individual safety [4]. Many efforts have been done to automatically detect and locate the abnormal events to avoid laborious and time-consuming work of manually recognizing them.

Existing AED methods can be mainly classified into three categories: *object-trajectory based methods*, *global-pattern based methods*, and *grid pattern based methods*. *Object-trajectory based methods* usually segment the crowd scene into different objects and then conduct objects tracking or identification. The trajectory of the object is the clue for abnormal events detection [5]. The abnormality of objects' trajectories are often evaluated by zone based analysis [6], fast matching algorithm [7], spatial-temporal path research [8], and deep learning algorithms [9]. Most of *global-pattern based methods* do not detect and track individuals in the scene separately. Instead, the goal of these methods is to extract low- or intermediate-level features from the video and analyzes the sequence as a whole [10]. Designing robust and descriptive features, which capture the unique properties of normal behavior, are quite commonly used. *Grid-pattern based methods* split frames into several blocks and analyze the pattern in blocks separately [11]. Sparse reconstruction cost [12], local features probabilistic framework [13], mixtures and temporal anomaly maps [14], low-rank and sparse decomposition [15], and joint sparsity model [16] are often utilized to evaluate the effectiveness of these methods.

While existing methods show significant performance, they highly depend on advanced computer vision technologies such

Manuscript received September 6, 2019; revised February 15, 2020 and June 26, 2020; accepted August 23, 2020. Date of publication September 14, 2020; date of current version October 7, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61906138, in part by the State Key Laboratory of Management and Control for Complex Systems Open Project under Grant 20190109, in part by the Shanghai Automotive Industry Sci-Tech Development Program under Grant 1838, in part by the European Union's Horizon 2020 Framework Programme for Research and Innovation (Human Brain Project SGA3) under Grant 945539, and in part by the Shanghai AI Innovation Development Program 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Karthik Nandakumar. *(Corresponding author: Guang Chen.)*

Guang Chen is with the Department of Automotive Engineering, Tongji University, Shanghai 200092, China, also with the Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich, 80333 Munich, Germany, and also with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: tj_autodrive@hotmail.com).

Peigen Liu, Zhengfa Liu, and Jinhu Dong are with the Department of Automotive Engineering, Tongji University, Shanghai 200092, China.

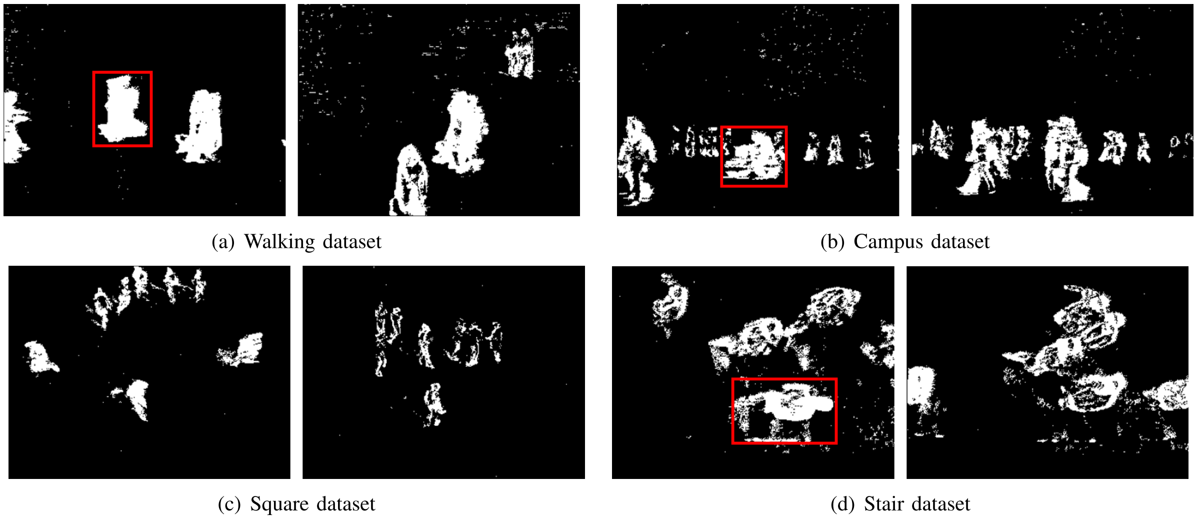
Huajin Tang is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China.

Lin Hong is with the College of Transportation, Shandong University of Science and Technology, Qingdao 266510, China.

Jörg Conradt is with the KTH Royal Institute of Technology, 114 28 Stockholm, Sweden.

Alois Knoll is with the Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich, 80333 Munich, Germany.

Digital Object Identifier 10.1109/TIFS.2020.3023791



(a) Walking dataset

(b) Campus dataset

(c) Square dataset

(d) Stair dataset

Fig. 1. Abnormal and normal event samples from NeuroAED dataset. In each sub-figure, left image contains abnormal event (marked with red square), right image only contains normal event. (a) Walking dataset, abnormal event: running, normal event: walking; (b) Campus dataset, abnormal event: bicycle, normal event: walking; (c) Square dataset, abnormal event: scattering, normal event: walking; (d) Stair dataset, abnormal event: wrong direction, normal event: walking down stairs.

as accurate object detection and tracking for *object trajectory based methods*, and robust feature descriptors extraction from images for *global pattern based methods*. In fact, the accuracy of object detection and tracking relies on the large amount of data for both online and offline model training. The heavy computation resources are always needed for image data processing. Further, when collecting data using standard frame-based cameras in static surveillance scenarios, it is unavoidable to record the redundant data from the background which are valueless for abnormal event detection.

To address these issues, we propose a new abnormal event detection system with neuromorphic vision sensor, which is named NeuroAED. Neuromorphic vision sensors such as the Dynamic Vision Sensor [17] are bio-inspired sensor that, in contrast to standard cameras, have *independent pixels* that output only intensity changes (called *events*[1]) asynchronously at the time they occurs. Comparing to the full images acquisition at a fixed frame rate of standard cameras, neuromorphic vision sensors have several advantages such as high dynamic range (140dB) and high temporal resolution (microseconds). Moreover, since changes of light intensity induced by moving object generating *events* in the scene, neuromorphic vision sensors are natural *motion detectors* and automatically filter out any redundant information such as static background in surveillance system [18]. As illustrated in Fig. 2(e) and Fig. 2(f).

Due to the unique principle of operation and unconventional output, new algorithms are developed in this work to exploit their capabilities in abnormal event detection. The major contributions of our work can be summarized in the following four aspects:

1) We develop a novel vision based abnormal event detection system that is different from most of the existing computer vision based methods.

2) To take the full advantages of the neuromorphic vision sensor, we develop a highly efficient event-based multiscale spatial-temporal (EMST) descriptor.

3) Considering the lack of a neuromorphic benchmark for the abnormal event detection, we record and publish the first neuromorphic vision based abnormal event detection dataset (called NeuroAED dataset). The dataset will be released,[2] and serves as a standard platform to shape the development of neuromorphic vision based abnormal event detection field.

4) Extensive experiments on NeuroAED dataset demonstrate that by using the proposed EMST descriptor, our system achieves great performance without relying on heavy computation of feature processing and complicate inference model.

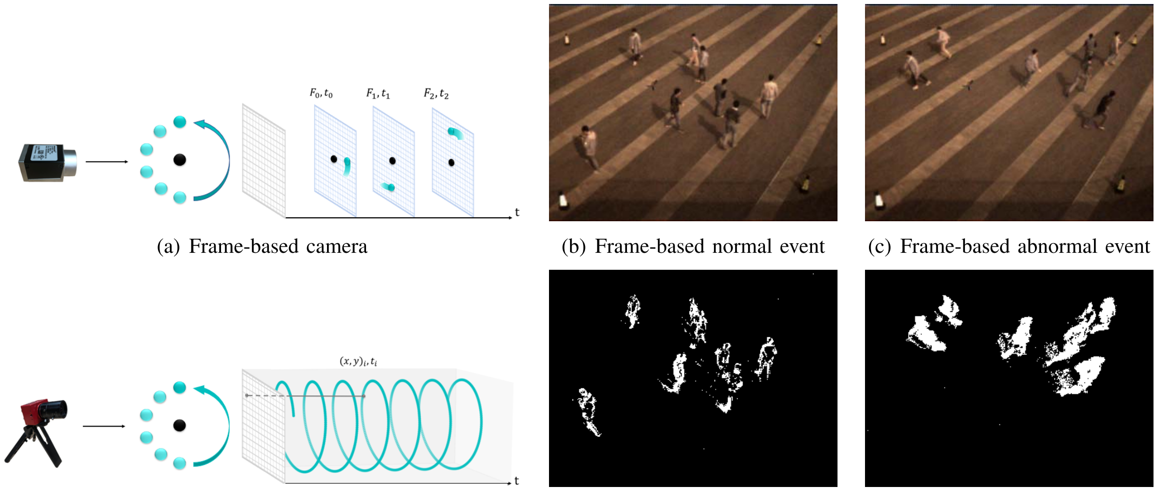
The rest of this paper is organized as follows: Section II gives a brief preliminary review of neuromorphic vision sensors. Section III presents the proposed NeuroAED system. Section IV describes our NeuroAED dataset. Section V shows experiment results. Section VI concludes this paper.

II. Preliminary

Neuromorphic vision sensors are bio-inspired sensors that work radically different from standard frame-based cameras. Instead of capturing images at a fixed rate, they response to pixel-level brightness changes asynchronously by generating a stream of *events*. In Fig. 2(a) and Fig. 2(d), the black ball is static while the green ball makes a fast circular motion around the black ball, the neuromorphic vision sensor and the standard frame-based camera are used to observe two balls simultaneously. Because the frame-based camera captures all pixel intensities at a fixed frame rate, the green ball appears

[1] To avoid any misunderstanding, the italic ***event*** represents the output data of a neuromorphic vision sensor. The normal **event** represents an incident happening on a scene such as abnormal event and normal event.

[2] The dataset and code are released at the link: https://github.com/ispc-lab/NeuroAED



(a) Frame-based camera (b) Frame-based normal event (c) Frame-based abnormal event

(d) Neuromorphic vision sensor (e) *Event*-based normal event (f) *Event*-based abnormal event

Fig. 2. Frame-based and *Event*-based normal and abnormal event. (a) Frame-based camera captures all pixel intensities at a fixed frame rate. (b) and (c): Frames captured by Active Pixel Sensor (APS) integrated in *DAVIS346* in NeuroAED dataset Square scene, where redundant background information are all recorded. (d) Neuromorphic vision sensor captures intensity changes caused by the moving objects asynchronously. (e) and (f): *Event* slices corresponding to (b) and (c), where static background information is automatically filtered out. Among them, (b) and (e) represent normal events, where the people are walking, (c) and (f) represent abnormal events, where people are running away.

as a blurred trajectory on the image due to the fast motion speed, as Fig. 2(a) shows. On the contrary, the neuromorphic vision sensor only captures brightness changes caused by the fast-moving green ball while information of stationary objects (black ball and background) are not recorded, as Fig. 2(d) shows.

The neuromorphic vision sensor has inherent advantages over motion detection, which inspires increasing interests and research efforts. Several newly neuromorphic vision datasets are developed recently. DvsGesture dataset [17] is the first gesture recognition dataset. Reference [19] proposes the first dataset for evaluating optical flow algorithms. For pose estimation and SLAM, [20] presents the world's first collection of datasets with a neuromorphic vision sensor for high-speed robotics. In [21], the first large-scale dataset dedicated to neuromorphic vision based intelligent driving is proposed.

Due to the unconventional output, algorithms and methods proposed based on frame-based cameras can not be directly extended to neuromorphic vision, and new algorithms are developed [18], [22]–[30]. Reference [22] integrates optical flow information computed at each *event* in a speed and direction coordinate frame to build motion-based feature for local corner detection and global gesture recognition. Reference [25] describes novel spatio-temporal features called time-surfaces for pattern recognition. Reference [23], [24] introduce *event*-driven categorization systems based on Spiking Neural Network. Reference [26] presents an neuromorphic vision based visual odometry algorithm, which leverages the outstanding properties of neuromorphic vision sensor to track fast motions while recovering a semidense 3D map of the environment. Reference [27] detects drowsiness driving using features extracted from *events* density. Reference [18] presents a deep neural network approach that unlocks the potential of neuromorphic vision sensor on a challenging motion-estimation task. Reference [28] presents an *event*-based visible light positioning (VLP) system, which can leave out the need for data association and traditional image processing methods. In [29], a novel *event*-based feature representation together with a new machine learning architecture is proposed and the first large real-world event-based dataset for object classification is released. In [30], a novel recurrent network is proposed to reconstruct videos from a stream of events, which aims to construct a bridge between neuromorphic vision algorithms and mature conventional vision algorithms. Reference [31] introduces the signal processing algorithms and applications for *event*-based neuromorphic vision in autonomous driving and various assistance systems. More emerging *event*-based algorithms are summered comprehensively in [32].

With the development of intelligent video surveillance systems, abnormal event detection research has achieved fruitful results due to the huge real life demands. Many of them focus on inventing algorithms of feature descriptors and modeling frameworks [6], [33]–[36]. In [6], an integrated pipeline that incorporates the output of object trajectory analysis and pixel-based analysis for abnormal behavior inference is proposed. This method enables to detect abnormal behaviors related to speed and direction of object trajectories, as well as complex behaviors related to finer motion of each object. In [34], the abnormal events are captured from a single static camera, a novel spatio-temporal feature descriptor, called histograms of optical flow orientation and magnitude and entropy is proposed based on optical flow information. In [35], a minimal path approach is used to model human trajectory behaviour for abnormal behavior detection. In the paper, the velocity and the orientation of the usual motion are considered to create a time surface on image plane, where each node/pixel shows the time needed to reach the pixel if the person behavior is normal and vice versa. In [36], a fully unsupervised dynamic sparse coding approach is proposed for detecting unusual events in videos. Based on online sparse reconstructibility of query

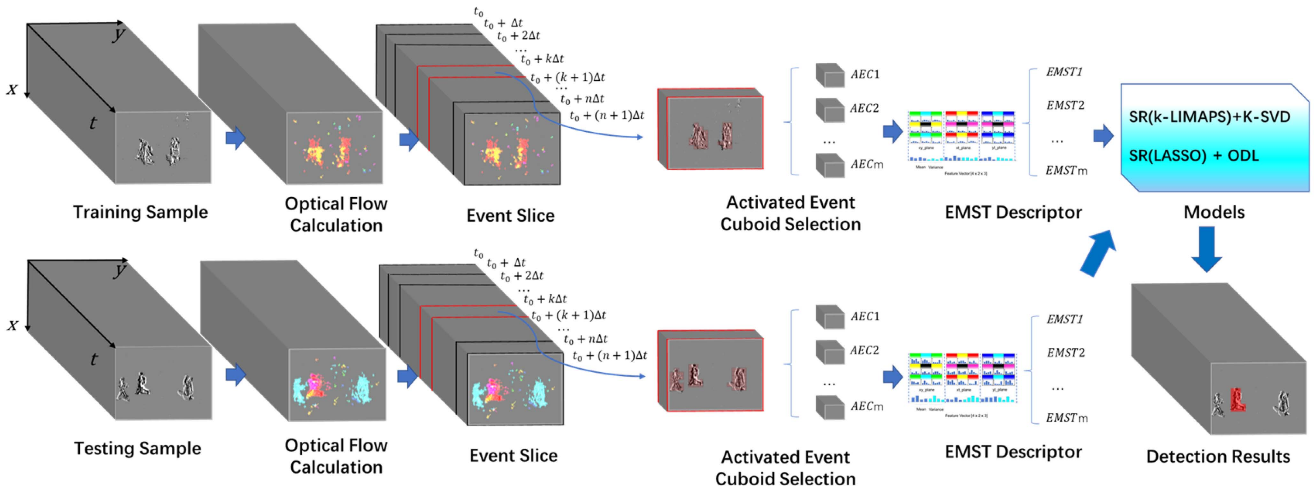


Fig. 3. Framework of the proposed NeuroAED system. We extract the optical flow information from training sample and select activated *event* cuboids based on the optical flow and *event* density to locate foreground. For each activated *event* cuboid, the proposed *event*-based multiscale spatio-temporal (EMST) descriptor is extracted and feed into models to learn the normal patterns. The trained models are used to identify descriptors of abnormal patterns extracted from the testing sample.

signals from an atomically learned event dictionary, which forms a sparse coding bases. In addition, [37] presents an event based algorithm to track vehicle and people. It's the first application of the neuromorphic vision sensor in surveillance system. Samsung recently launched an in-house monitoring system, the SmartThings Vision product [38], which can detect intruders and the falling in the house using a neuromorphic vision sensor.

III. Method

In this section, we present our NeuroAED system for abnormal event detection in detail. We divide our scheme into training and testing stages. An overview of our approach is illustrated in Fig. 3. Both stages consist of three main steps: optical flow extraction, activated *event* cuboid selection, and EMST descriptor generation. Our approach firstly divides the whole *event* stream into spaced *event* slices. For each *event* slice, activated *event* cuboids are selected. On the training stage, EMST descriptors are extracted from these cuboids which are kept as normal patterns. During testing, each activated *event* cuboid is identified by the trained model. The major technical contribution of proposed system focus on the construction of EMST descriptor, more specifically, the way of multiscale spatio-temporal encoding. First of all, in the pre-processing step, our method takes full advantages of dynamic sensitivity property of the neuromorphic vision sensor, and only capture foreground information by a simple scheme which is called activated cuboid. By using activated cuboid, it reduces large amount of data to be processed and improves the overall efficiency. Secondly, in order to guarantee the information contained in features that is rich enough, we integrate neighbour motion information of the activated cuboid on three different spatial-temporal planes and process them respectively. In this way, the features can not only record the motion characteristics of the cuboid itself, but also reveal the connections and differences with the adjacent region, which plays an important role in identifying abnormal cases. Thirdly, to handle the variety of different sizes of moving targets, the activated cuboid is expanded and shrunk respectively to obtain multi-scale features. Finally, features in different scales are concatenated.

A. Optical Flow Extraction

Neuromorphic sensor such as DVS sensor is a natural motion detector for moving objects on a scene. However, the direction and amplitude information can not be obtained directly from the events triggered by motion. We therefore choose optical flow to extract low-level features from the raw *event* stream as optical flow can characterize both the direction and amplitude information of an object movement. We assume that the moving direction and amplitude reflected by optical flow is the key to distinguish abnormal events from normal event in video surveillance. In this work, adaptive block-matching optical flow (ABMOF) presented in [39] is adopted. Fig. 4 shows its main principle, they use three time-slices, $t-2d$, $t-d$, t to accumulate previous and current *events* respectively, d is the slice duration. It should be noted that the event polarity is not used in the accumulating process, because it requires one bit of pixel memory to record and cannot improve accuracy significantly [39]. When a new *event* arrives, a reference block is generated in slice $t-d$ (red box in $t-d$), centered on the location of current *event*; then the best matching block is found in slice $t-2d$ based on the sum of absolute difference (red box in $t-2d$). The optical flow can be calculated using the offset of blocks and the time interval. To ensure the generated slices have enough features to match, the algorithm adopts a feedback control mechanism for adapting slice duration.

B. Activated Event Cuboid Selection

Extracting descriptors for each pixel of images from standard cameras is computationally expensive. It is encouraging

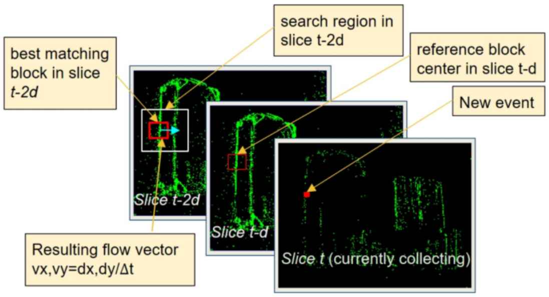

best matching block in slice t-2d
search region in slice t-2d
reference block center in slice t-d
New event
Slice t-2d
Slice t-d
Slice t (currently collecting)
Resulting flow vector vx,vy=dx,dy/Δt

Fig. 4. BMOF block matching [39]. It uses three time-slices to accumulate previous and current *events* respectively. When a new *event* arrives, a reference block is generated in slice $t - d$, centered on the location of current *event*; then the best matching block is found in slice $t - 2d$ based on the sum of absolute difference. The optical flow can be calculated using the offset of blocks and the time interval.

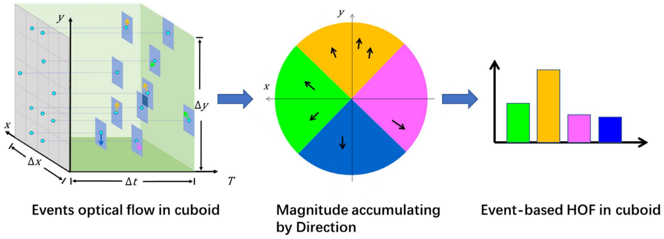

y
x
Δx
Δt
T
Δy
y
x
Events optical flow in cuboid
Magnitude accumulating by Direction
Event-based HOF in cuboid

Fig. 5. *Event*-based histogram of optical flow (eHOF) of the activated *event* cuboid. Left: the activated *event* cuboid with the size of $\Delta x \times \Delta y \times \Delta t$, blue squares represent *events* inside the cuboid, some of the events carry optical flow information, whose directions are distinguished by color. Middle: accumulating optical flow magnitude of *events* into o bins based on the corresponding direction. Right: eHOF feature of the cuboid (best viewed in color).

that DVS sensor does not acquire full images at a fixed frame rate but rather has independent pixels that output intensity changes. To this end, we design a very simple but highly efficient principle to select the so-called *activated event* cuboids which could be the candidate regions of abnormal events occur. Our approach firstly splits the *event* slice into $M \times N$ non-overlapping *event* cuboids with the size of $\Delta x \times \Delta y \times \Delta t$. Then, activated *event* cuboids are chose if the cuboids contain optical flow information and the *event* density are above a certain threshold. The threshold is set by extensive experiments by comparing the average value of the number of events in background and foreground cuboids in various scenarios. Because there is a huge margin, the threshold is simply set as the mean of event number of all the cuboids (half of them are foreground cuboids). As shown in Fig. 3, the selected activated *event* cuboids marked with dark red color cover the foreground of the scene quite well.

C. Event-Based Multiscale Spatio-Temporal Descriptors

With the optical flow information and the selected activated *event* cuboid, we aim to construct a robust descriptor for the abnormal event detection. Inspired by the histogram of optical flow (HOF) feature which is widely used in computer vision [33], our approach extracts the *event*-based histogram of optical flow feature (eHOF) for each activated *event* cuboid (See Fig. 5). We process *event*s inside each activated cuboid by accumulating their optical flow magnitude into o bins based on the corresponding direction to form eHOF feature, which is expressed as:

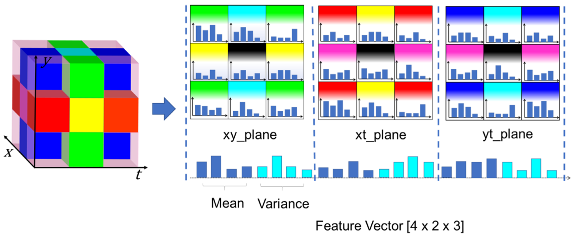


Fig. 6. Single scale spatio-temporal feature. Left: activated *event* cuboid and its neighbour cuboids. Centered on the activated *event* cuboid (black), neighbour cuboids of xy-*plane* (green), xt-*plane* (red) and yt-*plane* (blue) with identical size are selected, among them, yellow cuboids indicate the intersections between xy-*plane* and xt-*plane*, aqua green ones indicate intersections between xy-*plane* and yt-*plane*, purple ones indicate intersections between xt-*plane* and yt-*plane*. Right: in each plane, the average value and variance for each direction of the eHOF features of nine cuboids are calculated and concatenated to form the single scale spatio-temporal feature (best viewed in color).

$$H = (T_1, T_2, \ldots, T_o) \in R^{1\times o} \tag{1}$$

where T_i is the accumulated magnitude in the $i - th$ direction. We then choose eight adjacent cuboids around current activated *event* cuboid in each of the xy-*plane*, xt-*plane* and yt-*plane*. As shown in Fig. 6, activated *event* cuboid is colored in black. We take xy-*plane* as an example. Our method calculates eHOF feature for each neighbour cuboid, marked with green, yellow and aqua green. The neighbour yellow cuboids indicate the intersections between xy-*plane* and xt-*plane*, aqua green ones indicate intersections between xy-*plane* and yt-*plane*. In this way, the accumulated eHOF features of the c ($c = 9$, eight neighbour cuboids plus activated *event* cuboid) cuboids located in xy-*plane* are obtained:

$$P_{xy} = (H_{xy-1}, H_{xy-2}, \ldots, H_{xy-c}) \in R^{o\times c} \tag{2}$$

Then, the average value and variance for each direction are calculated:

$$\mu_{xy}^{(i)} = \frac{1}{c}\sum_{j=1}^{c} H_{xy-j}^{(i)} \in R^{1\times o} \tag{3}$$

$$\sigma_{xy}^{2(i)} = \frac{1}{c}\sum_{j=1}^{c} (H_{xy-j}^{(i)} - \mu_{xy}^{(i)})^2 \in R^{1\times o} \tag{4}$$

where, $i \in [1, o]$ represents the direction as mentioned before. We combine the $\mu_{xy}^{(i)}$ and $\sigma_{xy}^{2(i)}$ to form the description vector of xy-*plane* $F_{xy} = (\mu_{xy}^{(i)}, \sigma_{xy}^{2(i)}) \in R^{1\times 2o}$, which emphasizes on depicting spatial motion character in local region. Fig. 8 reveals how our descriptor uses spatial motion information to detect abnormal events. Fig. 8(a) shows two typical abnormal cases: (i) When an activated *event* cuboid and its neighbour cuboids most locate within the contour of a fast moving object or multiple fast moving objects with the same direction, their eHOFs will have larger magnitude on the motion direction, which results in the correspond average value of F_{xy} becoming larger. (ii) When activated *event* cuboid locate near the contour

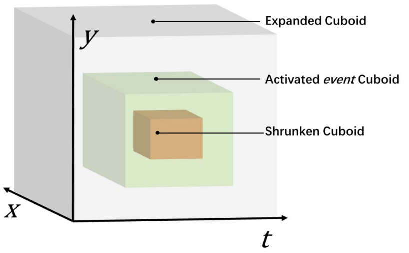


Fig. 7. Multiscale spatio-temporal cuboids. The green cuboid represents the original activated *event* cuboid. The gray and orange cuboids are obtained by expanding and shrinking in both space and time dimensions (best viewed in color).

of fast moving objects, the eHOFs of neighbour cuboids inside contour will have larger magnitude on the moving direction, in contrast, the eHOFs of neighbour cuboids outside the contour, *i.e.* lie in background region or slow moving objects, will have smaller magnitudes, which results in the correspond variance becoming large. By comparing with two typical normal cases given in Fig. 8(b), we can find out that, either the mean or variance of abnormal F_{xy} will be larger than the normal ones, which is helpful to identify abnormal events. Note that, here, we just use one direction to elaborate. In case of multiple directions, differences in features can be reflected in each direction.

To describe the motion characteristic more specifically, F_{xt} and F_{yt} are calculated in the same manner. Different from F_{xy} that focuses on describing motion characteristic with respect to spatial space, these two descriptors pay attention to extracting motion features in time sequence. Their variance values contain the velocity change in time dimension, which correspond to the acceleration value that serve as an important evaluation indicator. Then, description vectors of three planes are combined to form the single scale *event*-based spatio-temporal descriptor $F_s = (F_{xy}, F_{xt}, F_{yt})$.

In order to deal with the scale variation of the moving objects, each activated *event* cuboid in *event* slice is resized into three scales, as shown in Fig. 7. The green cuboid represents the original activated *event* cuboid. The gray and orange cuboids are obtained by expanding and shrinking in both space and time dimensions. Then, single scale spatio-temporal features are extracted from the resized cuboids respectively, and the final *event*-based multiscale spatio-temporal descriptor is acquired by concatenating them $F_m = (F_s, F_{exp}, F_{shr})$.

D. Abnormal Event Detection

The main idea of the abnormal event detection is to learn the normal event model and classify the testing inputs as normal or abnormal according to the trained model. The abnormal event is detected based on whether it deviates from or violates the normal patterns. In this work, sparse representation (SR) model is used to model normal event patterns for abnormal event detection, which has been thoroughly studied and extensively tested in [12], [15], [16]. It consists of four parts, including Dictionary learning, Sparse encoding, Iterative updating, and Abnormal measurement. First, we construct a dictionary of features from the input features. Following, we employ the dictionary to obtain the sparse coding results of the input features. Furthermore, we calculate the reconstruction loss to update the dictionary which could be used for sparse coding again. By alternative updating between the dictionary learning and sparse encoding, we can obtain sparse encoding results of the input features. After that, we apply the rarity similarity between the input features to predict the anomaly score of each testing input. In the following, we will describe these parts separately.

1) Dictionary Learning: The goal of dictionary learning is to learn an effective dictionary from an basis feature set to obtain the sparse coding results. Assuming that the input features is $\mathbf{X} \in \mathbb{R}^{m\times n}$, we want to learn an overcomplete dictionary $D \in R^{m\times d}$ $(m < d)$, so that the feature x could be described by sparse linear combinations of the atoms with the coefficients β. The dictionary learning can be formulated as:

$$\min_{\beta} \|X - D\beta\|_2^2 \quad s.t. \quad \|\beta\|_p \le s \tag{5}$$

where $\|\beta\|_p$ is the penalty or regularization constraint of parameter s $(\le n)$ to produce sparse representation.

Generally, D and β can be alternatively optimized. When D is fixed, the problem is called sparse coding. And vice versa, if β is fixed, the problem is called dictionary learning. In this, in order to demonstrate the efficiency of the designed EMST descriptor, we utilize K-SVD [40] and ODL [41] algorithms to construct the dictionary from the input EMST features, respectively. For a given scene in video stream, a set of training input EMST features can be described as $X = \{F_{m1}, F_{m2}, \ldots, F_{mn}\}$, where m is the dimension of the EMST descriptor, n is the number of the training normal cuboids.

2) Sparse Coding: With the learned dictionary, we employ sparse encoding algorithm to obtain the sparse representation α of each EMST descriptor. In this work, we employ k-LIMAPS [42] and LASSO algorithms for the sparse coding task, respectively. After sparse coding, the input features X are represented as $R(X) = \left\{R_{F_{m1}}, R_{F_{m2}}, \ldots, R_{F_{mn}}\right\}$.

3) Iterative Updating: For the first iteration, we use the initialization dictionary, which is built by randomly sampling from the input features, to obtain the sparse representations of the input features. Then the sparse representations are used to calculate reconstruction errors for updating the dictionary by minimizing the reconstruction errors. And for the following iterations, the updated dictionary is utilized for sparse coding again. Through jointly optimizing between the dictionary learning and sparse coding, we obtain effective dictionary and sufficiently sparse representations of the input features. In this work, two model, SR (k-LIMPAS)+K-SVD and SR (LASSO)+ODL, are built.

4) Abnormal Measurement: Based on the learned dictionary and sparse representations, we capture the inherent structures and patterns of the input data to detect anomalies. Specifically, we measure the abnormality of an event according to the rarity similarity between the input features. The atoms indicate the basis vectors in the over-complete dictionary D. Each atom could be regarded as an attribute or a feature. If two inputs

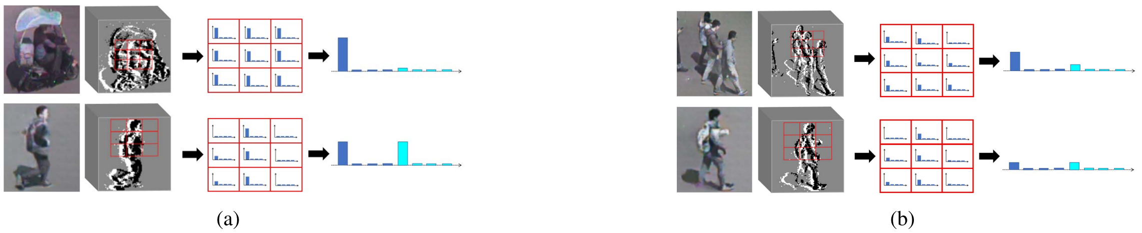


Fig. 8. Spatial motion characteristic embedded in single scale spatio-temporal feature for *xy-plane* (a) Abnormal events. Up: an activated *event* cuboid and its neighbour cuboids most locate within the contour of a fast moving object, their eHOFs will have larger magnitude on the motion direction, which results in the correspond average value of F_{xy} becoming large; Bottom: when activated *event* cuboid locate near the contour of fast moving objects, the eHOFs of neighbour cuboids inside contour will have larger magnitude on the moving direction, in contrast, the eHOFs of neighbour cuboids outside the contour, *i.e.* lie in background region or slow moving objects, will have small magnitudes, which results in the correspond variance becoming large. (b) Normal events. Either the mean or variance of normal events is smaller, no matter where the cuboids locate at.

share some specific atoms and the corresponding coefficients are close, it's obvious that they could be similar to each other in terms of motion patterns (speed or direction). Therefore, the rarity similarity can be used to predict the anomaly scores of each testing input. We use Mahalanobis distance to measure similarity in this work. In the training phase, we construct the statistical distribution of the sparse representations of all training inputs, where the mean values and the covariance matrix of the sparse representations are computed. In the testing phase, we calculate the Mahalanobis distance between the sparse representation of each testing input and the statistical distribution for anomaly prediction based on a certain detection threshold θ, given by

$$\text{Label}\left(X_{\text{test}}\right) = \begin{cases} \text{normal} & M\left(R\left(X_{\text{test}}\right)\right) < \theta \\ \text{abnormal} & M\left(R\left(X_{\text{test}}\right)\right) \geq \theta \end{cases} \tag{6}$$

where $M(R(X_{test}))$ is the Mahalanobis distance, and $R(X_{test})$ is the sparse representations of the testing feature. The θ is obtained through training, which represents maximum distance of all training inputs.

IV. NeuroAED Dataset

Frame-based computer vision algorithms with standard cameras are in rapid development partially due to the widely accepted datasets, which allow direct comparison between algorithms. Considering the important role of datasets playing in the development of abnormal event detection system and the lack of a neuromorphic vision based abnormal event dataset, we build the first neuromorphic vision dataset dedicated to the abnormal event detection, which is named NeuroAED dataset. The NeuroAED dataset comprises 152 samples of four different indoor and outdoor scenarios, and is split into four sub-dataset: Walking, Campus, Square and Stair dataset. And each dataset contains two slice sequences: training samples and testing samples. The training samples only contain normal events, while testing samples are both normal and abnormal events. For each slice sample of the NeuroAED dataset, the groundtruth annotation of a binary flag indicating normal or abnormal events occur is provided. With the exception of the Square dataset, the manually generated pixel-level binary masks are contained in each slice sample, which identify the abnormal events regions. More details refer to IV-B and Table I.

A. Dataset Recording

The NeuroAED datasets are acquired with a stationary neuromorphic vision sensor *DAVIS346* with a 346 × 260 pixel resolution mounted on the top of a retractable tripod with a maximum elongation of five meters, in which a pan-tilt is used to adjust camera angle for covering the entire region of interest. For Walking and Campus dataset, the data are recorded from walkways of a college campus with pedestrian movement parallel to the camera plane, and the abnormal events occur naturally, e.g. bike or motorcycle. For Square and Stair dataset, volunteers are required to stage for assembling the data. The video footage recorded from each dataset is chopped into various clips varying from 10 seconds to 24 seconds, referring to Table I for details.

B. Dataset Description

1) Walking Dataset: The Walking dataset is recorded on a walking street on a sunny day. The data are captured through overlooking the walking street. It contains 30 training samples and 28 testing samples. Most of the samples have a rather sparse crowd density. The duration of each sample is around 8-20 seconds. The normal events only contain people walking. The abnormal events are due to either anomalous pedestrian motion patterns (running) or non-pedestrian moving objects (bike, motorcycle), as Fig. 1(a) shows. The pixel-level binary masks are provided for evaluating the anomaly localization.

2) Campus Dataset: The Campus dataset is collected in a campus walkway on a cloudy day. And different from the Walking dataset, it is captured in a horizontal view. The crowd density in the walkway varies from sparse to very crowded. It contains 30 training samples and 30 testing samples. Each sample has a duration of 5-14 seconds. The definition of both the normal and abnormal events are same as the Walking dataset, as the Fig. 1(b) shows.

3) Square Dataset: The Square data is only for slice level abnormal event detection inspired by the UMN dataset, and recorded at a square. It consists of 12 training samples and 6 testing samples with 7-10 seconds and 10-16 seconds time interval, respectively. Each of the testing samples starts with an initial part of normal event and ends with abnormal behavior sequences. The normal events refer to people walking around in the square, and the abnormal events are people suddenly scattering to different directions, as Fig. 1(c) shows.

TABLE I
DESCRIPTION OF THE NEUROAED DATASET

Scenario	Number of samples		T(s)		Ground Truth		Abnormal Events	Normal Events
	Testing	Training	Testing	Training	SL	PL		
Walking dataset	28	30	8-20	10-19	+	+	running, bicycle, motorcycle	walking
Campus dataset	30	30	5-10	6-14	+	+	bicycle,motorcycle	walking
Square dataset	6	12	7-10	10-16	+	-	scattering	walking
Stair dataset	6	10	14-20	13-24	+	+	running, wrong direction	walking down stairs

T: Duration of each sample, SL: Slice-Level, PL: Pixel-Level.

4) Stair Dataset: The stair dataset is collected in an indoor stair, and the lighting condition is poor in which 10 training samples and 6 testing samples are recorded. Each one lasts about 13-24 seconds. The normal events are people walking down stairs. The abnormal events mainly include wrong direction events and running events, as Fig. 1(d) shows. Being same with the Walking dataset and Campus dataset, the Stair dataset has pixel-level groundtruth annotation to identify the abnormal events regions.

V. EXPERIMENTS AND ANALYSIS

In this section, extensive experiments are conducted on our NeuroAED dataset to evaluate the performance of the EMST descriptor for abnormal event detection, and demonstrate its effectiveness by comparing the state of the art approaches.

A. Evaluation Metrics

Two commonly used measurements are adopted to evaluate the performance of abnormal event detection [43]: *Slice-Level* and *Pixel-level*. All measurements consider the matching between the evaluated result and the ground-truth.

1) Slice-Level: If one or more cuboids are detected as abnormal cuboids in a testing *event* slice, it is labelled as an abnormal slice. If the ground truth of this slice is abnormal, it is a True Positive (TP). Otherwise, it is a False Positive (FP).

2) Pixel-Level: In pixel-level measurement, a detected abnormal slice is TP if more than 40% truly abnormal pixels are detected. A normal slice is FP as long as one pixel is detected as abnormal. Pixel-level measurment emphasizes the correct detection of abnormal objects.

3) Computational Cost: The computational cost is represented by the ratio between the time spent in detecting abnormal events (including the process of optical flow extraction, descriptors construction and abnormal events identifying) on all testing samples and the total duration of all testing samples. The lower the computational cost is, the better the efficiency of the method is. We implement our algorithm using c++ for feature generation and using python for model training on a PC with 8 GB RAM and 1.60 GHz Intel i5 8250u processor.[3]

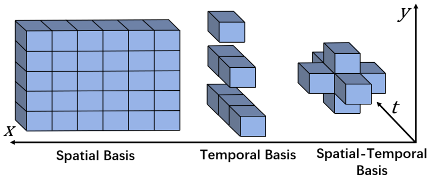


Fig. 9. Various kinds of basis descriptors. Left: Spatial Basis descriptor(SB), covers whole slice with non-overlapping cuboids and the eHOF features of all cuboids are concatenated to form the feature of the basis. Middle: Temporal Basis descriptor, when a activated *event* cuboid is selected, based on which, temporal basis with different time span are created, the eHOF features of cuboids are concatenated to form the feature of the basis. Right: Spatial-Temporal Basis descriptor(STB), based on activated *event* cuboid, neighbour cuboids in time and spatial dimension are created to form the STB descriptor, and the eHOF features of cuboids are concatenated to form the basis feature.

B. Experiment Setup

1) Baseline Methods: In order to evaluate our method more comprehensive, inspired by the frame-based descriptors in [12], we build *event*-based spatial and temporal basis descriptors to detect abnormal events, as shown in Fig. 9. The spatial basis covers whole slice and is suitable for detecting global abnormal events. For detecting local abnormal events, when a activate cuboid is selected, based on which, spatial-temporal basis and temporal basis with different time span are created. For each unit of the descriptor, the eHOFs are calculated and concatenated to form the features of basis. Next, we use both SR(k-LIMAPS)+K-SVD and SR(LASSO)+ODL models to learn and identify feature vectors generated by them.

2) Parameter Setup: Since the proposed descriptor can adapt to various scales of moving objects, the EMST parameters can be set to identical values in different scenarios. We split the input video into non-overlapping cuboids with the spatial size of 18×14 pixels and $100ms$ duration, the *event* number threshold α for activated *event* cuboid selection is set to 200. As for the spatial basis of baseline descriptor, the whole slice is split into 6×5 units with $100ms$ duration.

C. Experiment Results

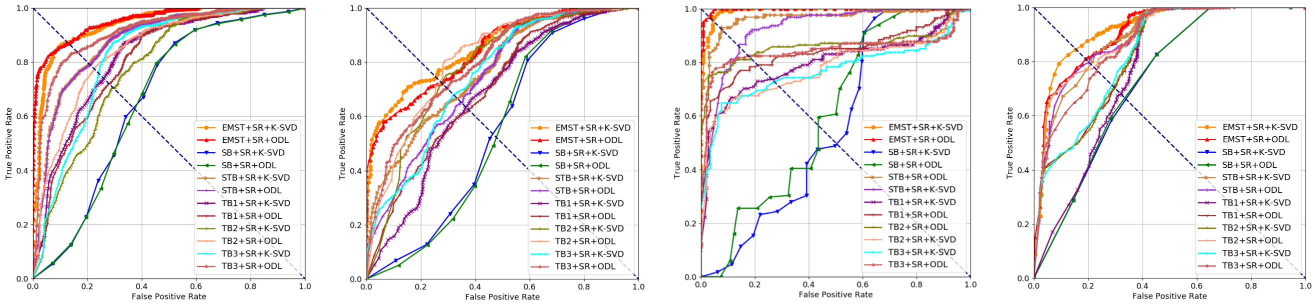
1) Experiments on Walking Dataset: In this section, the detection results and the analyses on Walking dataset are given. Selected visualized results are shown in Fig. 12 and the detected abnormal events are marked with red rectangles. The abnormal events include running, bicycle and motorcycle. Slice-level Receiver Operator Characteristic (ROC) curves are shown in Fig. 10 and pixel-level ROC curves are

[3] It needs to be noted that the optical flow extracting method [39] is integrated in jAER software, which can not provide the exactly computing time. The extracting time is provided by the author of [39] upon our request, which is calculated based on their implementation on the FPGA platform.

TABLE II
COMPARISON OF SLICE-LEVEL AND PIXEL-LEVEL ON ALL DATASETS

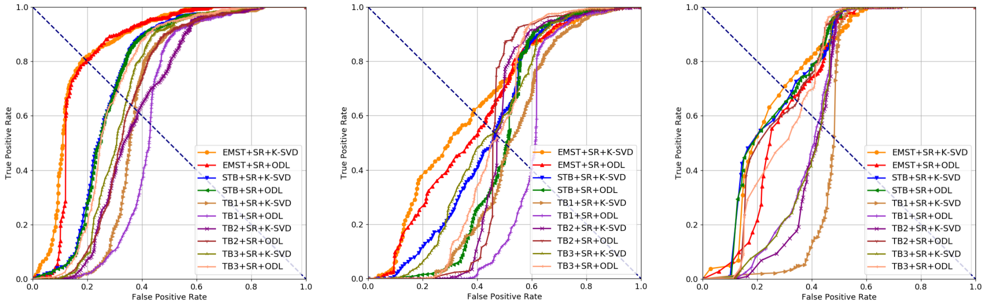
METHODS	Walking dataset				Campus dataset				Square dataset		Stair dataset			
	Slice-Level		Pixel-Level		Slice-Level		Pixel-Level		Slice-Level		Slice-Level		Pixel-Level	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
EMST+SR+K-SVD	95.5	13.0	**87.9**	18.7	**85.7**	**25.5**	**65.7**	**38.5**	98.3	4.3	**92.0**	**15.3**	74.9	32.2
EMST+SR+ODL	**95.8**	**12.5**	87.6	**17.3**	84.7	27.7	63.2	42.8	**99.7**	**3.5**	90.6	19.1	71.2	33.5
SB+SR+K-SVD	66.5	34.0	-	-	55.4	45.3	-	-	56.5	49.6	72.9	28.6	-	-
SB+SR+ODL	66.2	34.9	-	-	55.1	46.6	-	-	60.7	43.5	73.1	28.2	-	-
STB+SR+K-SVD	87.8	21.1	73.3	30	76.8	33.5	56.9	46.5	92.8	14.2	88.8	21.6	75.2	32.7
STB+SR+ODL	87.8	20.8	73.1	30.1	74.3	35.1	51.2	50.4	96.8	7.9	89.1	20.1	**75.4**	**30.2**
TB1+SR+K-SVD	82.3	25.9	62.5	37.5	67.8	36.4	47.1	51.5	79.8	27.1	77.3	32.2	54.2	48.3
TB1+SR+ODL	80.7	28.2	58.2	43.1	70.3	37.6	40.6	60.1	84.1	21.5	82.7	29.7	62.0	42.5
TB2+SR+K-SVD	77.6	29.7	63.2	38.6	78.1	27.6	52.3	46.2	78.0	29.2	82.7	29.4	60.2	43.7
TB2+SR+ODL	84.5	22.7	65.9	36.3	81.2	26.8	52.7	45.8	86.2	18.1	89.2	22.3	73.4	35.0
TB3+SR+K-SVD	82.3	24.1	68.2	34.3	75.2	31.9	56.4	46.5	77.0	27.8	83.2	28.6	61.9	43.6
TB3+SR+ODL	91.0	16.2	71.9	31.2	80.2	29.7	54.0	47.4	84.7	17.4	87.3	23.0	70.8	36.5

[1] Descriptor: EMST (event based multiscale spatio-temporal); SB (spatial basis); STB (spatial-temporal basis); TBn(temporal basis with n units).
[2] Model: SR+K-SVD (sparse representation model and K-SVD dictionary learning method); SR+ODL (sparse representation model and online dictionary learning method).



(a) (b) (c) (d)

Fig. 10. Slice-level ROC curves. (a) Walking Dataset; (b) Campus Dataset; (c) Square Dataset; (d) Stair Dataset.



(a) (b) (c)

Fig. 11. Pixel-level ROC curves. (a) Walking Dataset; (b) Campus Dataset; (c) Stair Dataset.

shown in Fig. 11. Based on these ROC curves, Area Under Curve (AUC) and Equal Error Rate (EER) are computed and listed in Table II. In addition, the computational cost of various methods on the Walking dataset are listed in Table III.

For the slice-level comparison in Table II, the AUC of our work is 95.8% and the EER is 12.5%. For the pixel-level comparison in Table II, the AUC of our work is 87.9% and the EER is 18.7%. Our results outperform the baselines methods, because in our work, the EMST descriptor catch more spatial and temporal connection information by considering multiple spatio-temporal scales, which makes the representations of activated *event* cuboids more robust. For the computational cost reported in Table III, we can see that, our method encode and process rich information at the price of slightly larger computational cost. However, considering that our implementation is not optimized for high-speed processing, it is possible to further reduce the computational cost of the proposed method.

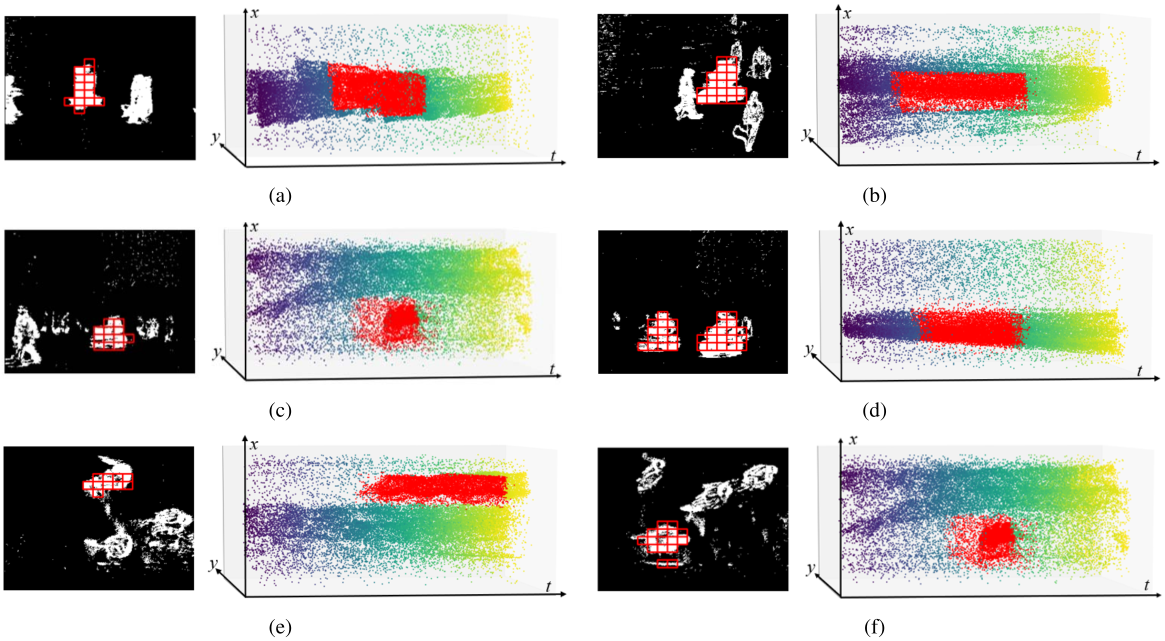


Fig. 12. Demonstration of the abnormal event detection results in NeuroAED dataset. Fig. (a) and (b) are samples from Walking dataset. Fig. (c) and (d) are samples from Campus dataset. Fig. (e) and (f) are samples from Stair dataset. In each sub-figure, left-side image shows the red-rectangle marked abnormal events on top of the accumulated *event* slice and right-side image shows the red-color marked abnormal events in the spatial-temporal space of the raw *event* slice.

TABLE III

COMPARISON OF THE COMPUTATIONAL COST ON THE WALKING DATASET. THE LOWER THE COMPUTATIONAL COST IS, THE BETTER THE EFFICIENCY OF THE METHOD IS

METHODS	Computational Cost
EMST+SR+K-SVD	0.433
EMST+SR+ODL	0.355
SB+SR+K-SVD	0.220
SB+SR+ODL	**0.209**
STB +SR+K-SVD	0.310
STB +SR+ODL	0.221
TB1+SR+K-SVD	0.241
TB1+SR+ODL	0.282
TB2+SR+K-SVD	0.251
TB2+SR+ODL	0.239
TB3+SR+K-SVD	0.271
TB3+SR+ODL	0.313

[1] Descriptor: EMST (event based multiscale spatio-temporal); SB (spatial basis); STB (spatial-temporal basis); TBn(temporal basis with n units).

[2] Model: SR+K-SVD (sparse representation model and K-SVD dictionary learning method); SR+ODL (sparse representation model and online dictionary learning method).

2) Experiments on Campus Dataset: In this section, the detection results and the analyses on Campus dataset are given. Selected visualized results are shown in Fig. 12 and the detected abnormal events are marked with red rectangles. The abnormal events include bicycle and motorcycle. Slice-level ROC curves are shown in Fig. 10 and pixel-level ROC curves are shown in Fig. 11. Based on these ROC curves, AUC and EER are computed and listed in Table II.

For the slice-level comparison in Table II, the AUC of our work is 85.7% and the EER is 25.5%. For the pixel-level comparison in Table II, the AUC of our work is 65.7% and the EER is 38.5%. Comparing with Walking dataset, Campus dataset has a lower AUC and EER. As both datasets have very similar abnormal and normal events, we think that the differences may come from the sensor setup. For the campus dataset, data are recorded in a horizontal view which an abnormal object is likely to be blocked by a walking pedestrian, so the abnormal event such as moving bicycle is not detected. On the other hand, due to the wider range of depth in horizontal view, objects with same ground truth speed can correspond to quite different optical flow information at different depth, which would influence the training and detection results. However, for the walking campus, the data are captured with a *DAVIS 346* mounted on top of a retractable tripod which enjoy a broad field of view and has limited range of depth, which can make abnormal objects appear on the scene all the time and optical flow for objects with different speed are more distinguishable.

3) Experiments on Square Dataset: In this section, the detection results and the analyses on Square dataset are given. The abnormal event is pedestrian scattering to different directions. Slice-level ROC curves are shown in Fig. 10 and pixel-level ROC curves are not available. Based on these ROC curves, AUC and EER are computed and listed in Table II. It is not surprising that our approach achieves 99% on this dataset as the abnormal event such as crowd scattering differ from normal pedestrian walking in several aspects such as the velocity and moving direction of the pedestrians.

4) Experiments on Stair Dataset: In this section, the detection results and the analyses on stair dataset are given. Selected visualized results are shown in Fig. 12 and the detected abnormal events are marked with red rectangles. The abnormal events include running and wrong direction. Slice-level ROC curves are shown in Fig. 10 and pixel-level ROC curves are shown in Fig. 11. Based on these ROC curves, AUC and EER are computed and listed in Table II.

For the slice-level comparison in Table II, the AUC of our work is 92.0% and the EER is 15.3% that outperforms other work. For the pixel-level comparison in Table II, the AUC of our work is 74.9% and the EER is 32.2%, the baseline method spatial-temporal basis achieves best performance with 75.4% AUC and 30.2% EER. This is because the floor and stair handrail are all reflective material, which cause the appearance of many shadows. And EMST descriptor takes multi-scale neighbour region into account that will be influenced by shadows seriously and can not locate the anomalies accurately. However, the STB is a single scale descriptor, which encodes less neighbour information and is less affected by shadows compared with our work. On the other hand, comparing with various temporal basis, STB catches more spatial and temporal connection information, which makes it outperform other work.

5) Parameter Analysis: In this section, we analyze the effect of various parameters on detection performance using the Walking dataset, including both detection accuracy and computational cost. The parameters include the *event* number threshold α for the activated *event* cuboid selection, the number of orientations for eHOF and the scale of cuboid. Besides, the multiscale spatio-temporal character of EMST is also analyzed by comparing with single scale spatio-temporal feature, *i.e.* do not change spatial and temporal scale in the process of generating features. The analysis results are shown in Table IV and Table V.

We analyze the *event* number threshold α first. From Table IV, it can be found that with the increase of α, the detection performance is getting slightly better until $\alpha = 200$, then remains unchanged basically, this is because some noise *events* are introduced into the construction process of EMST when smaller α adopted, and reduce the detection accuracy, but this effect is negligible since the robustness of the descriptor. The noise would be filtered clearly enough if α is greater than 200, thus keeps the performance more steady. On the other hand, the number of features required to be generated and processed are decreasing as α rises, which improves the computational efficiency.

Then, the impact of the number of orientations of eHOF feature on detection performance is analyzed, we test eHOF features with two and eight orientations respectively and compare the results with four orientations eHOF (used in proposed EMST). It can be seen that the number of orientations has little effect on detection performance. The two orientations eHOF is slightly better than others, which may be caused by clearly distinguishable motions in this dataset. In addition, the computational cost of two orientations method is minimal, eight and four orientations methods have roughly equal computational cost.

TABLE IV

COMPARISON OF DETECTION PERFORMANCE AND COMPUTATION COST FOR DIFFERENT *EVENT* NUMBER THRESHOLD α ON THE WALKING DATASET

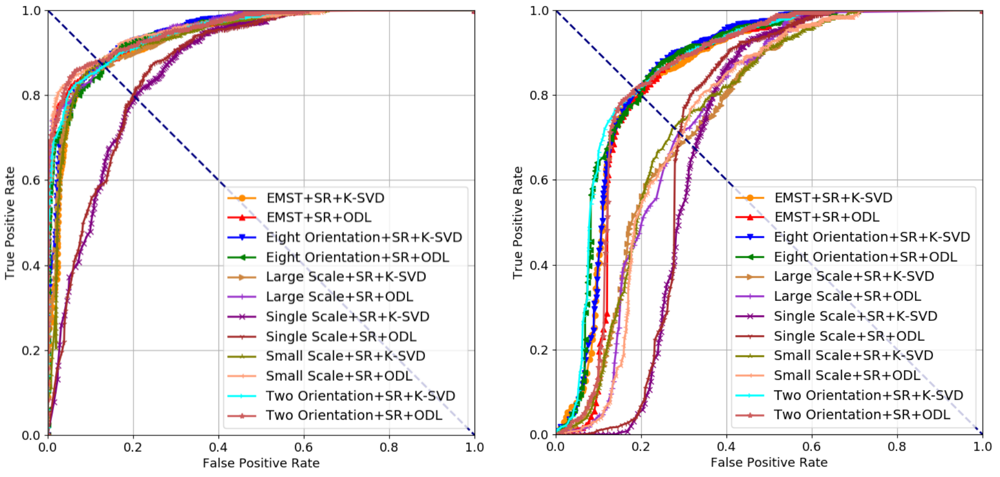
METHODS	α	Slice-Level		Pixel-Level		Computational Cost
		AUC	EER	AUC	EER	
EMST+SR+K-SVD	100	95.2	13.3	87.5	18.3	0.478
	150	95.3	13.7	87.5	18.7	0.456
	200	95.5	13.0	87.9	18.7	0.433
	250	95.4	13.2	87.7	18.8	0.415
	300	95.6	13.1	88.1	17.9	0.403
	350	95.5	13.2	**88.2**	18.5	0.384
	400	95.6	12.9	**88.2**	18.0	0.373
EMST+SR+ODL	100	95.7	**12.4**	87.4	18.3	0.387
	150	95.7	12.8	87.3	17.8	0.367
	200	95.8	12.5	87.6	**17.3**	0.355
	250	95.7	13.0	87.3	17.5	0.344
	300	95.8	13.4	87.7	17.7	0.335
	350	**95.9**	13.3	87.9	17.4	0.324
	400	**95.9**	13.1	87.7	18.5	**0.319**

[1] EMST (event based multiscale spatio-temporal);
[2] SR+K-SVD (sparse representation model and K-SVD dictionary learning method); SR+ODL (sparse representation model and online dictionary learning method).

Further, various scales of the cuboid are also analyzed, including cuboids of 22 × 16 × 100 (Large Scale), cuboids of 18 × 14 × 100 (EMST) and cuboids of 14 × 12 × 100 (Small Scale). It needs to be noted that the *event* number threshold α is chosen based on the volume of cuboid, *i.e.* $\alpha = [(22 \times 16 \times 100)/(18 \times 14 \times 100)] \times 200 = 280$ for large scale and 130 for small scale. Although small scale features achieves the best performance in slice-level among all of them, the corresponding pixel-level performance and computational cost is less satisfactory. On the other hand, the computational cost of large scale methods is the smallest, but its pixel-level performance is worse than others.

Furthermore, by comparing the performance of EMST and single scale feature, we can see that the multiscale spatial-temporal character of proposed descriptor has a significant enhancement on the detection performance, which, on the other hand, increases the computational cost.

6) Comparison With the Frame-Based Methods: In this section, we compare our method with two frame-based approaches [44] [45] for anomaly detection on Walking dataset. Two kinds of frames are adopted in this work (See Fig. 14). The first one is the standard frame directly captured by the APS of DAVIS346. The second one is the reconstructed frame based on the events with the method proposed in [30]. The first frame-based method [44] proposes a discriminative framework for abnormal event detection. By following their approach, we compute gradient-based features for 10 × 10x5 (rows x columns x frames) spatio-temporal sub-units in testing videos and each sub-unit is represented by a 100 dimensional vector after Principal Component Analysis (PCA) and normalization. The second frame-based method [45] presents a combination of spatial feature extractor and temporal sequence ConvLSTM to detect anomalies. They transform the detection of abnormal event into spatiotemporal sequence outlier detection problem, and build an end-to-end model by incorporating convolutional feature extractor in both spatial and temporal



(a) (b)

Fig. 13. Comparison of slice-level (a) and pixel-level (b) detection results for various EMST parameters on the walking dataset.

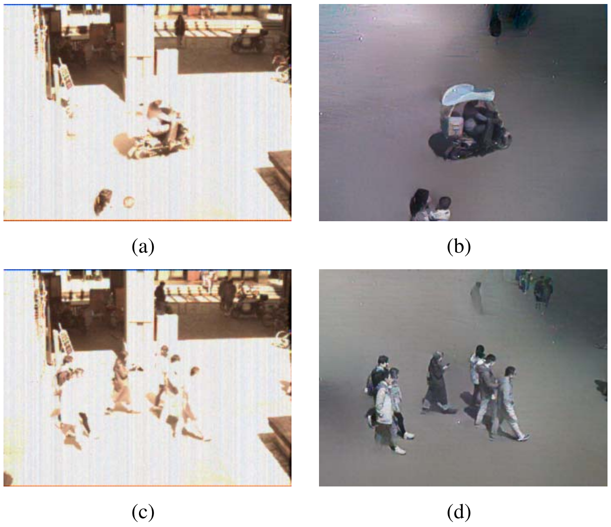
TABLE V

COMPARISON OF DETECTION PERFORMANCE AND COMPUTATION COST FOR VARIOUS PARAMETERS ON THE WALKING DATASET

METHODS	Slice-Level		Pixel-Level		Computational Cost
	AUC	EER	AUC	EER	
EMST+SR+K-SVD	95.5	13.0	87.9	18.7	0.433
EMST+SR+ODL	95.8	12.5	87.6	17.3	0.355
Two Orientation+SR+K-SVD	96.0	12.8	88.2	18.5	0.380
Two Orientation+SR+ODL	96.2	12.1	88.0	18.6	0.341
Eight Orientation+SR+K-SVD	94.9	12.9	85.8	18.6	0.412
Eight Orientation+SR+ODL	95.2	12.9	86.7	19.6	0.361
Small Scale+SR+K-SVD	93.8	12.3	78.1	28.0	0.502
Small Scale+SR+ODL	96.4	12.0	74.9	29.5	0.425
Large Scale+SR+K-SVD	94.2	13.3	75.7	31.1	0.340
Large Scale+SR+ODL	95.6	12.1	75.2	29.0	0.299
Single Scale+SR+K-SVD	86.9	20.5	69.1	32.7	0.270
Single Scale+SR+ODL	87.2	19.9	70.7	28.9	**0.242**

[1] Descriptor: EMST (event based multiscale spatio-temporal); Two Orientation, Eight Orientation (the number of orientations of eHOF of EMST); Small Scale, Large scale (the scale of cuboid of EMST are 14×12×100 and 22×16×100 respectively); Single scale (EMST without expanded and shrunken cuboid).

[2] Model: SR+K-SVD (sparse representation model and K-SVD dictionary learning method); SR+ODL (sparse representation model and online dictionary learning method).



(a) (b) (c) (d)

Fig. 14. The sample frames used by frame-based methods in the Walking dataset. (a) and (c) are captured by the APS of DAVIS346. (b) and (d) are reconstructed using the method [30].

space into the encoding-decoding structure. The quantitative comparisons in terms of AUC, EER and computation cost are shown in Table VI. It needs to be noted that the method [44] is tested under Matlab 2018a environment on the same PC with our method. The method [45] is tested with a server with 64GB RAM, 2.50GHz Intel E5 processor and 1080Ti GPU. It gets the best computational cost at the price of advanced computation hardware. Our proposed approach outperforms the frame-based methods while it is also more efficient than the frame-based method [44].

7) Discussion: It is interesting to see that by relying on simple optical flow information, our work achieves an average AUC of 93.3% at the slice level and 76.3% at the pixel level. This indicates that our EMST descriptor is robust to different scenes. Although our system can detect most of the abnormal events, there are still failed cases. Because the *event* stream from neuromorphic vision sensor contains limited appearance information, it is challenging to detect abnormal objects which has similar pattern of velocity and motion with normal objects. Some failure cases are shown in Fig. 15, (a) is a motorcycle moving very slowly in the Walking dataset, it is difficult for our method to detect it as our method does not consider either the geometry or the appearance of the moving objects. Fig. 15(b) is a pedestrian moving fast near the neuromorphic sensor, which our approach fails to distinguish from the other pedestrians, because the proposed EMST is a motion descriptor that relies on optical flow information which will become large when object moving fast near the sensor and cause false positive. Part of the reason is that the neuromorphic visions sensor DAVIS346 has a pretty low image resolution (346 × 260 pixel array). Fig. 15(c)

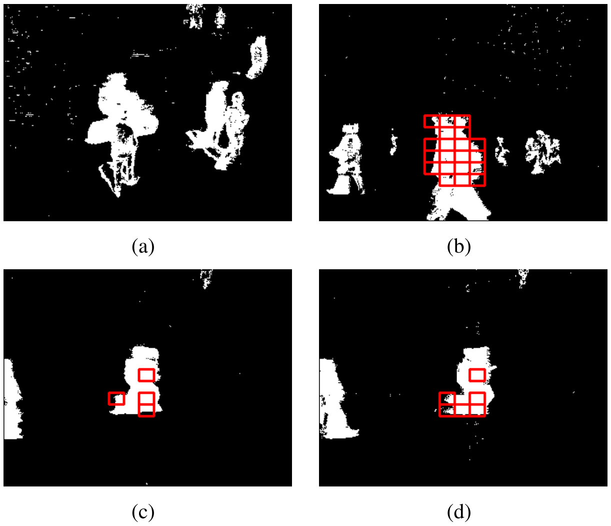

(a) (b) (c) (d)

Fig. 15. The failure cases in Walking and Campus dataset. (a) is a motorcycle with low speed, which is missed by our method. (b) is a pedestrian moving fast near the neuromorphic sensor, which causes false positive. (c) and (d) are two continuous *event* slices, showing a person running through the street, which are detected incompletely and discontinuously.

TABLE VI
COMPARISON WITH FRAME BASED METHODS ON THE WALKING DATASET

METHODS	Slice/Frame-Level		Computational Cost
	AUC	EER	
EMST+SR+K-SVD	95.5	13.0	0.433
EMST+SR+ODL	**95.8**	**12.5**	0.355
RC Allison *et al.* [44]	80.7	27.7	1.832
APS Allison *et al.* [44]	79.4	28.3	1.761
RC Yong *et al.* [45]	84.3	23.1	0.223*
APS Yong *et al.* [45]	81.2	26.5	**0.214***

EMST: event based multiscale spatio-temporal; SR+K-SVD: sparse representation model and K-SVD dictionary learning method; SR+ODL: sparse representation model and online dictionary learning method;
RC: frames are reconstructed from events using [30]; APS: frames are captured by APS of DAVIS346 directly;
* The computation cost is calculated on a server with 1080Ti GPU. Others are calculated with a PC.

and 15(d) are two continuous *event* slices, showing a person running through the street. Both cases have the defects of incomplete detection, which is caused by the incorrect optical flow information extracted. Since the construction of EMST descriptor is highly dependent on optical flow, it is unlikely to prevent such failures from happening. In addition, we can find that the detection of (c) and (d) is discontinuous, this is because the abnormal event is detected by a short *event* slice, which makes the detection discontinuous. It only evaluates the abnormal events based on the current activated cuboid (although our works expand the scale of the cuboid). If an abnormal object occupies several cuboids in several *event* slices, only the detected abnormal cuboid will be marked. Nevertheless, we aim at taking this opportunity to understand how the abnormal detection could benefit from the natural response of neuromorphic vision to motion and their inherent data redundancy reduction. However, it will be an interesting direction for future work to exploit the fusion of *event* stream and RGB images for abnormal event detection.

Our future work will focus on these drawbacks and limitations and exploit better solutions for these challenges.

VI. Conclusion

In this work, we show a novel neuromorphic vision-based abnormal event detection system for visual surveillance, named NeuroAED. In NeuroAED system, we firstly design a simple but efficient method based on the sparse *events* to select activated *event* cuboids, which locate the interesting regions from the foreground fast and accurately. We then design a novel *event*-based multiscale spatio-temporal descriptor to extract features from the activated *event* cuboids for the abnormal event detection. Additionally, we build the NeuroAED dataset, the first public dataset dedicated to abnormal event detection. Experiments are conducted based on this dataset and demonstrate the high efficiency and accuracy.

Our NeuroAED system addresses the issues such as expensive computation resources and large amounts of data storage, which are suffered by the traditional visual surveillance systems based on RGB cameras. The neuromorphic vision sensor used in this work automatically filters out any temporally-redundant information, thus there is no need for background video data storage and background subtraction processing. The NeuroAED system is designed to exploit the natural motion detection characteristics of the neuromorphic vision sensor, which is able to detect abnormal events efficiently showing great potential of on-line and cloud computation and storage for future visual surveillance.

Acknowledgment

The authors would like to thank all the volunteers who helped in the collecting process of the dataset.

References

[1] M. V. Anees and S. G. Kumar, "Deep learning framework for density estimation of crowd videos," in *Proc. 8th Int. Symp. Embedded Comput. Syst. Design (ISED)*, Dec. 2018, pp. 16–20.
[2] C. Hu *et al.*, "Building an intelligent video and image analysis evaluation platform for public security," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
[3] R. Kaviani, P. Ahmadi, and I. Gholampour, "Incorporating fully sparse topic models for abnormality detection in traffic videos," in *Proc. 4th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2014, pp. 586–591.
[4] B.-W. Chen, C.-Y. Chen, and J.-F. Wang, "Smart homecare surveillance system: Behavior identification based on state-transition support vector machines and sound directivity pattern analysis," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 43, no. 6, pp. 1279–1289, Nov. 2013.
[5] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Trajectory-based surveillance analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1985–1997, Jul. 2019.
[6] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Bremond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017.
[7] L. Patino, J. Ferryman, and C. Beleznai, "Abnormal behaviour detection on queue analysis from stereo cameras," in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2015, pp. 1–6.
[8] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3488–3496.

[9] D. Tran, J. Yuan, and D. Forsyth, "Video event detection: From subvolume localization to spatiotemporal path search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 404–416, Feb. 2014.

[10] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition-a review," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 865–878, Nov. 2012.

[11] H. Sfar and A. Bouzeghoub, "Activity recognition for anomalous situations detection," *IRBM*, vol. 39, no. 6, pp. 400–406, Dec. 2018.

[12] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognit.*, vol. 46, no. 7, pp. 1851–1864, Jul. 2013.

[13] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2112–2119.

[14] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[15] X. Cui, Y. Tian, L. Weng, and Y. Yang, "Anomaly detection in hyperspectral imagery based on low-rank and sparse decomposition," in *Proc. 5th Int. Conf. Graphic Image Process. (ICGIP)*, vol. 9069, Jan. 2014, Art. no. 90690R.

[16] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive sparse representations for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 631–645, Apr. 2014.

[17] A. Amir *et al.*, "A low power, fully event-based gesture recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7388–7397.

[18] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5419–5427.

[19] B. Rueckauer and T. Delbruck, "Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor," *Frontiers Neurosci.*, vol. 10, p. 176, Apr. 2016.

[20] E. Mueggler, H. Rebecq, G. Gallego, T. Delbrück, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM," *Int. J. Robot. Res.*, vol. 36, no. 2, pp. 142–149, Feb. 2017.

[21] J. Binas, D. Neil, S. Liu, and T. Delbrück, "DDD17: End-to-end DAVIS driving dataset," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017.

[22] C. Xavier, M. Jean-Matthieu, B. Sébastien, and R. B. Benosman, "A motion-based feature for event-based pattern recognition," *Frontiers Neurosci.*, vol. 10, p. 594, Jan. 2017.

[23] B. Zhao, R. Ding, S. Chen, B. Linares-Barranco, and H. Tang, "Feedforward categorization on AER motion events using cortex-like features in a spiking neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 1963–1978, Sep. 2015.

[24] R. Xiao, H. Tang, Y. Ma, R. Yan, and G. Orchard, "An event-driven categorization model for aer image sensors using multispike encoding and learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3649–3657, Sep. 2020.

[25] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "HOTS: A hierarchy of event-based time-surfaces for pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1346–1359, Jul. 2017.

[26] H. Rebecq, T. Horstschaefer, G. Gallego, and D. Scaramuzza, "EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real time," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 593–600, Apr. 2017.

[27] G. Chen, L. Hong, J. Dong, P. Liu, J. Conradt, and A. Knoll, "EDDD: Event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor," *IEEE Sensors J.*, vol. 20, no. 11, pp. 6170–6181, Jun. 2020.

[28] G. Chen *et al.*, "A novel visible light positioning system with event-based neuromorphic vision sensor," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10211–10219, Sep. 2020.

[29] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "HATS: Histograms of averaged time surfaces for robust event-based object classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1731–1740.

[30] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3857–3866.

[31] G. Chen, H. Cao, J. Conradt, H. Tang, F. Röhrbein, and A. Knoll, "Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 34–49, Jul. 2020.

[32] G. Gallego, T. Delbruck, G. M. Orchard, C. Bartolozzi, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 10, 2020.

[33] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 988–998, Jun. 2014.

[34] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017.

[35] B. Cancela, A. Iglesias, M. Ortega, and M. G. Penedo, "Unsupervised trajectory modelling using temporal information via minimal paths," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2553–2560.

[36] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3313–3320.

[37] M. Litzenberger *et al.*, "Embedded vision system for real-time object tracking using an asynchronous transient vision sensor," in *Proc. IEEE 12th Digit. Signal Process. Workshop 4th IEEE Signal Process. Edu. Workshop*, Sep. 2006, pp. 173–178.

[38] P. K. J. Park *et al.*, "Low-latency interactive sensing for machine vision," in *IEDM Tech. Dig.*, Dec. 2019, pp. 10.6.1–10.6.4.

[39] M. Liu and T. Delbruck, "Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018.

[40] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[41] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 2, pp. 19–60, 2010.

[42] A. Adamo, G. Grossi, R. Lanzarotti, and J. Lin, "Sparse decomposition by iterating lipschitzian-type mappings," *Theor. Comput. Sci.*, vol. 664, pp. 12–28, Feb. 2017.

[43] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981.

[44] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 334–349.

[45] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.* Cham, Switzerland: Springer, 2017, pp. 189–196.