Guang Chen, Hu Cao, Jörg Conradt, Huajin Tang,
Florian Röhrbein, and Alois Knoll

Event-Based Neuromorphic Vision for Autonomous Driving

A paradigm shift for bio-inspired visual sensing and perception

©ISTOCKPHOTO.COM/OONAL

Digital Object Identifier 10.1109/MSP.2020.2985815
Date of current version: 26 June 2020

As a bio-inspired and emerging sensor, an event-based neuromorphic vision sensor has a different working principle compared to the standard frame-based cameras, which leads to promising properties of low energy consumption, low latency, high dynamic range (HDR), and high temporal resolution. It poses a paradigm shift to sense and perceive the environment by capturing local pixel-level light intensity changes and producing asynchronous event streams. Advanced technologies for the visual sensing system of autonomous vehicles from standard computer vision to event-based neuromorphic vision have been developed. In this tutorial-like article, a comprehensive review of the emerging technology is given. First, the course of the development of the neuromorphic vision sensor that is derived from the understanding of biological retina is introduced. The signal processing techniques for event noise processing and event data representation are then discussed. Next, the signal processing algorithms and applications for event-based neuromorphic vision in autonomous driving and various assistance systems are reviewed. Finally, challenges and future research directions are pointed out. It is expected that this article will serve as a starting point for new researchers and engineers in the autonomous driving field and provide a bird's-eye view to both neuromorphic vision and autonomous driving research communities.

Introduction

Over the past few decades, the rapid development of electronics, information technologies, and artificial intelligence have made great progress in artificial visual sensing and perception systems. For example, the vision system of an autonomous vehicle becomes more intelligent by using deep learning technology. However, it still has some shortcomings compared with biological counterparts, such as the human and animal visual systems. Even small insects, such as bees, outperform the most advanced artificial vision systems such as high-quality cameras in routine functions, including real-time sensing and processing, low-latency motion control, and so on. More importantly, such biological neural systems can well perform

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

tasks with small energy consumption. In fact, biological neural systems usually consist of a large number of relatively simple elements. They operate in a massively parallel principle, which is different from the most common type of vision sensors such as CMOS cameras. Thus, some researchers and engineers have tried to mimic the working principles of the biological visual systems and come up with a new artificial visual system.

Recently, the developments of material technologies, lithographic processes, very large-scale integration (VLSI) design techniques, neuroscience, and neuromorphic technologies have enabled the novel conception and fabrication of bio-inspired visual sensors and processors. These new sensors and processors provide different methods to sense and perceive the world. The event-based neuromorphic vision sensor is such a bio-inspired vision sensor mimicking biological retina from both the system level and element level; it poses a paradigm shift in the way of visual information acquisition, processing, and modeling. The dynamic vision sensor (DVS) proposed by the group of Tobi Delbruck [1] is the first practicable event-based neuromorphic vision sensor based on the biological principle. DVS captures the per-pixel brightness changes (called *events*) asynchronously instead of measuring the absolute brightness of all pixels at constant rate, resulting in promising properties compared to standard frame-based cameras, such as low power consumption and low latency (in the order of microsecond), HDR (120 dB), and high temporal resolution [2]. Thus, an alternative visual sensing and perception system for autonomous vehicles is provided in challenging scenarios that state-of-the-art standard frame-based cameras cannot well perform [3], [4], such as high-speed scenes of the autonomous highway driving, low latency of motion control, and low power consumption of the vehicle onboard system.

It is well known in the research of autonomous driving that radar, lidar, ultrasound, and cameras form the backbone of sensor systems of the autonomous vehicle [5]–[7]. These sensors acquire the visual data as a sequence of snapshots recorded at discrete time stamps; therefore, visual information is compressed and quantized at a predefined frame rate. Consequently, a problem that is often known from the signal processing domain (undersampling) arises due to the timescale of motions in the observed scenes and the frame-rate of the recording camera. Things occurring between the adjacent frames, along with the consequent information, would get lost. Generally, the advanced algorithms with multiple-sensor fusion are usually developed to compensate single-sensor shortcomings in demanding applications such as highly piloted driving systems with low-latency motion control and visual feedback loops. Rather than solving this problem from an algorithmic perspective, it is better to explore alternative methods from a novel sensing perspective, such as event-based neuromorphic vision sensors. This results in providing great value for promoting subsequent tasks to become more robust, accurate, and complementary together with advanced algorithm development.

As an emerging sensing technology, the algorithms and applications of event-based neuromorphic vision are in the preliminary stage.

As an emerging sensing technology, the algorithms and applications of event-based neuromorphic vision are in the preliminary stage. Some works have been summarized in [8]. Unlike [8], this article aims to provide a thorough overview of the event-based neuromorphic vision for autonomous driving, from a signal processing perspective with a focus on visual perception algorithms and applications (see Figure 1). Specifically, the introduction starts from the operation principle of this bio-inspired neuromorphic vision sensor; then, the unique advantages of the sensor and its connection with the perception system of autonomous vehicles are discussed. Taking these promising properties into consideration, the signal processing techniques about event noise processing, event data representation, and meaningful event-based neuromorphic vision algorithms of given autonomous driving tasks are illustrated. Afterward, the works of event-based neuromorphic vision that are dedicated to specific applications in autonomous driving are reviewed. Finally, we address the problems remaining to be tackled and the directions for future research.

Bio-inspired vision

A biological retina

The retina of vertebrates, such as humans, is a highly developed multilayer neural system consisting of light sensitive cells which contain millions of photoreceptors. It is the place where the acquisition and preprocessing of the visual information happen. As shown in Figure 2(c), the retina has three primary layers including the photoreceptor layer, the outer plexiform layer, and the inner plexiform layer.

The photoreceptor layer consists of light-sensitive cells that convert incoming light into electrical signals and drive the horizontal cells and bipolar cells in the outer plexiform layer. There are two major types of bipolar cells: ON- and OFF-bipolar cells. The ON- and OFF-bipolar cells are responsible for coding the bright and dark spatial-temporal contrast changes, respectively. Particularly, the firing rate of the ON-bipolar cells will increase while the OFF-bipolar cells will no longer generate spikes if the illumination is increasing. This, in turn, increases the firing rate of OFF-bipolar cells in the case of illumination decreasing (such as getting darker). In the absence of a light stimulus, both cells generate few random spikes. This phenomenon is achieved by comparing the photoreceptor's signals with the spatial-temporal values, which are determined by the mean value of the horizontal cells, facilitating the connection between photoreceptors and bipolar cells laterally. In the outer plexiform layer, the ON- and OFF-bipolar cells synapse onto the amacrine cells and ON- and OFF- ganglion cells in the inner plexiform layer. The amacrine cells mediate signal transmission between bipolar cells and ganglion cells. The ganglion cells carry information along with different parallel pathways in the retina, which is conveyed to the visual cortex. Thus, the retina is responsible for converting spatial-temporal

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

illumination changes into pulses, which is transmitted to the visual cortex via the optic nerve.

Silicon retina

Silicon retinas are visuals that model the biological retina and follow neurobiological principles. Pioneers of silicon retinas are Mahowald and Mead, who introduced their silicon VLSI retina in 1991 [9]. This kind of sensor is equipped with adaptable photoreceptors and a chip with a 2D hexagonal grid of pixels. It replicates parts of cell types of biological retinas, including the photoreceptors, bipolar cells, and horizontal cells. Therefore, this kind of sensor represents merely the photoreceptor layer and the outer plexiform layer. Later, Zaghloul and Boahen built the Parvo-Magno retina, which is superior to the silicon VLSI retina, by modeling five retina layers.

Despite the promising structure, many of the early silicon retinas originate from the biological sciences community and are mainly used to demonstrate neurobiological models and theories without considering real-world applications. Recently, an increasing amount of effort from Tobi Delbruck's team has been put into the development of practicable silicon retina DVS based on biological principles [1]. In Figure 2, the three-layer model of a human retina [Figure 2(c)], and corresponding DVS pixel circuitry [Figure 2(a)] are presented. Typical signals of the pixel circuits are displayed in Figure 2(b). The upper trace denotes a voltage waveform at the node $v_{\log}$, which tracks

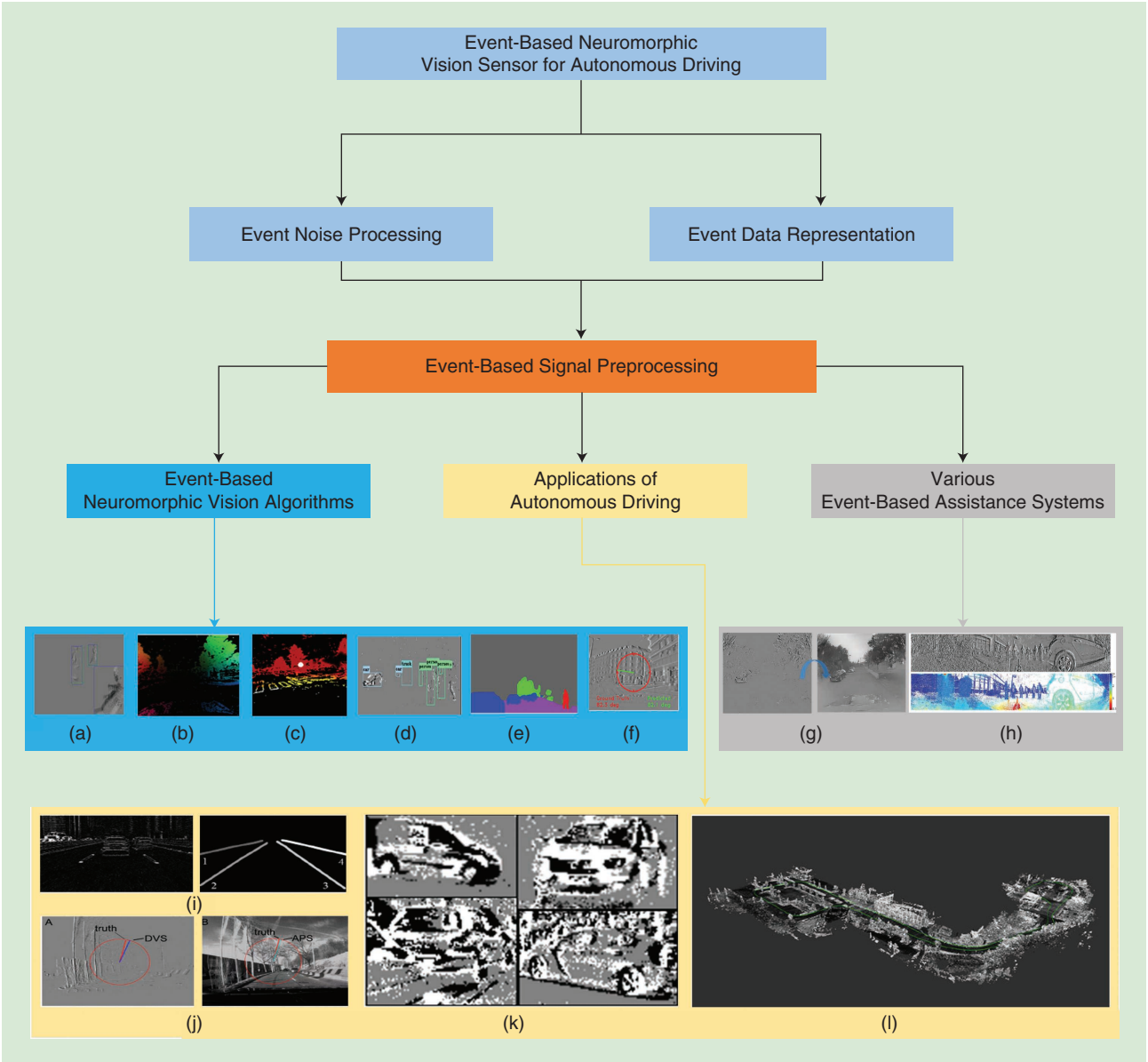

FIGURE 1. An overview of event-based neuromorphic vision sensors for autonomous driving, with representative examples for emerging systems and applications: (a) tracking (adapted from [22]), (b) optical flow (adapted from [29]), (c) depth estimation (adapted from [29]), (d) object detection (adapted from [52]), (e) semantic segmentation (adapted from [35]), (f) steering prediction (adapted from [3]), (g) image reconstruction (IR) (adapted from [45]), (h) panoramic stereo vision (adapted from [47]), (i) DET data set (adapted from [16]), (j) DDD17 data set [18] (adapted from [3]), (k) N-Cars data set (adapted from [17]), and (l) MVSEC data set (adapted from [4]).

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

the photocurrent through the photoreceptor layer circuit. The outer plexiform layer circuit responds with spike events (v_{diff}) of different polarities to positive and negative changes of the photocurrent. Spikes are transported to the next processing stage by the inner plexiform layer circuit. A large number of log-intensity changes are encoded in the events. Figure 2(d) illustrates the accumulated events including ON event (illumination increased) and OFF event (illumination decreased) that are drawn as white and black dots.

Today's representatives of silicon retinas are mainly from pioneers Tobi Delbruck and Christoph Posch and represent a compromise between biological and technical aspects. In their development, one prominent challenge posed is usually regarded as a wiring problem, indicating that each pixel of the silicon retina needs its own cable, which is impossible for chip wiring. A key technique for the solution, named *address event representation* (*AER*) was originally from the Caltech group of Carver Mead; it is used as an event-controlled and asynchronous point-to-point communication protocol for prototypes of the silicon retina.

As illustrated in Figure 3, the basic functionality of AER is implemented by an address encoder (AE), an address decoder (AD), and a digital bus. All neurons and pixels could transmit the time-coded information on the same line because the digital bus implements a multiplex strategy. The AE of the sending chip generates a unique binary address for each neuron or pixel in case of a change. The bus transmits the address at high speed to the receiver chip. Then, the AD determines the position and generates a spike on the receiver neuron. Event streams are employed in AER to communicate among chips. An event is a tuple (x, y, t, p); x and y are pixel addresses; t is the time stamp; and p represents the polarity. The polarity indicates the increase and decrease in the lighting intensity, corresponding to an ON event and OFF event, respectively.

This article focuses mainly on the first practically usable silicon retina, the DVS, which follows the natural, frame-free, and event-driven approach that triggers a plethora of research in event-based neuromorphic vision and autonomous driving. [A recent approach by Tobi Delbruck is the so-called dynamic and active pixel vision sensor (DAVIS) that combines dynamic and static visual information into a single pixel.] The DVS pixel models a simplified three-layer biological retina by mimicking the information flow of the photoreceptor–bipolar–ganglion cells (see Figure 2). Pixels operate independently and attach special importance to the temporal development of the local lighting intensity. The DVS pixel would automatically trigger an event (either ON event or OFF event) when the relative change in intensity exceeds the threshold. Therefore, the working principle of the DVS is fundamentally different from the

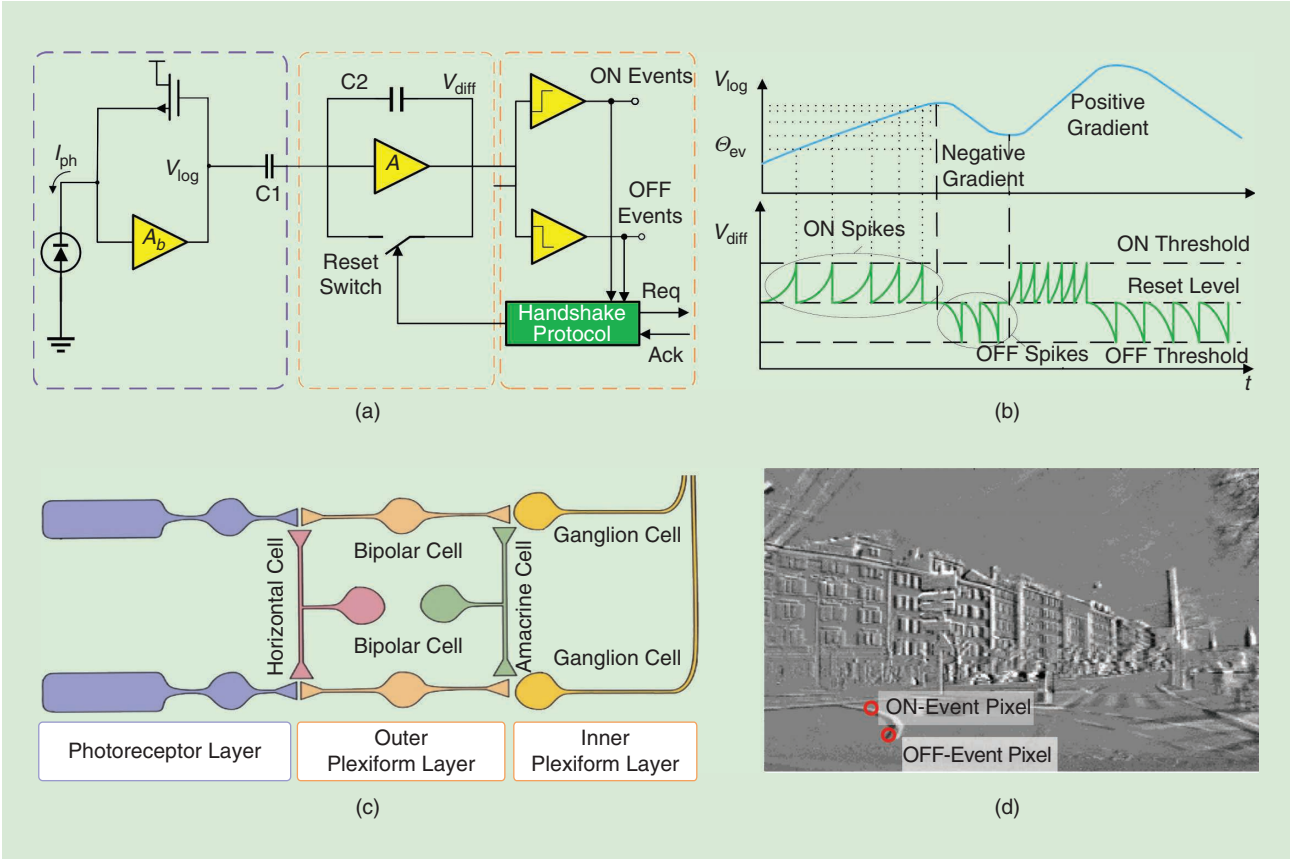

FIGURE 2. A practicable silicon retina DVS based on biological principles (adapted from [53]): (a) DVS pixel circuitry, (b) typical signals of the pixel circuits, (c) a three-layer model of a human retina, and (d) the accumulated events from a DVS. The accumulated event map has ON event (illumination increased) and OFF event (illumination decreased) drawn as white and black dots.

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

frame-based camera. There are three key properties of biological vision that are kept in this silicon retina: the relative illumination change, the sparse event data, and the separate output channels (ON/OFF). The major consequence of the DVS is that the acquisition of visual information is no longer controlled by any form of external timing signals such as frame clock or shutter, while the pixel itself controls its own visual information individually and autonomously.

Advantages of bio-inspired vision sensors

Due to the fundamentally different working principle and the mimicking of the biological retina, the event-based neuromorphic vision sensors have several advantages over standard frame-based cameras.

- *Energy-friendly properties*: Since event-based neuromorphic vision sensor transmits only events and autonomously filter redundant data, power is only used to process active pixels (e.g., the events triggered by illumination changes). Particularly, an energy-friendly sensor is more important than advanced algorithms for the onboard computers and devices in autonomous vehicles.
- *Low latency*: There is no need for the global exposure of the frame because each pixel works independently. Ideally, the minimal latency is 10 μs. The low-latency control of the autonomous vehicle is highly dependent on the perception systems. A low-latency perception system such as an object-detection system based on an event-based neuromorphic vision sensor would save lots of time in avoiding obstacles for the control systems.
- *HDR*: The event-based neuromorphic vision sensor such as DVS has an HDR (120 dB), which far exceeds that of the frame-based cameras (60 dB). Event-based neuromorphic vision sensors such as the DVS can simultaneously adapt to very dark and bright stimuli ensuring a highly robust perception system even in a light-changing scene such as an autonomous vehicle driving through a tunnel.
- *Microsecond resolution*: The brightness changes can be captured quickly in analog circuity. With a 1-MHz clock, events can be detected and time-stamped with microsecond resolution. Considering the fast response requirement of the controller in autonomous vehicles in emergency driving scenes, this property is quite useful in autonomous driving.
- *No motion blur*: In the high-speed driving scenario, the motion blur problem occurs when the motion of the moving objects is beyond the sampling frequency of the frame-based camera; this may cause the failure of the perception system. An event-based neuromorphic vision sensor can capture dynamic motion precisely with no motion blur; it is of great value to autonomous driving community.

Event noise processing

The preprocessing of the raw data is essential for extracting meaningful information for sensor systems. An event-based neuromorphic vision sensor not only captures the change in the light intensity caused by moving objects, it also generates some noise activities due to the movements of background objects and the sensor noise such as temporal noise and junction leakage currents [10]–[12]. As shown in Figure 4, the event noise processing technique is responsible for excluding the event noises from the event stream. Two commonly used methods in the literature, namely the spatial-temporal correlation filter and the motion consistency filter, are illustrated as follows.

Spatial-temporal correlation filter

For a newly incoming event $e_i = (x_i, y_i, t_i, p_i)$, the spatial-temporal filter searches the most recent neighborhood event around the current pixel location (x_i, y_i) within a distance D. The incoming event would be regarded as a nonnoise event if the time difference meets:

$$t_i - t_n < d_t, \tag{1}$$

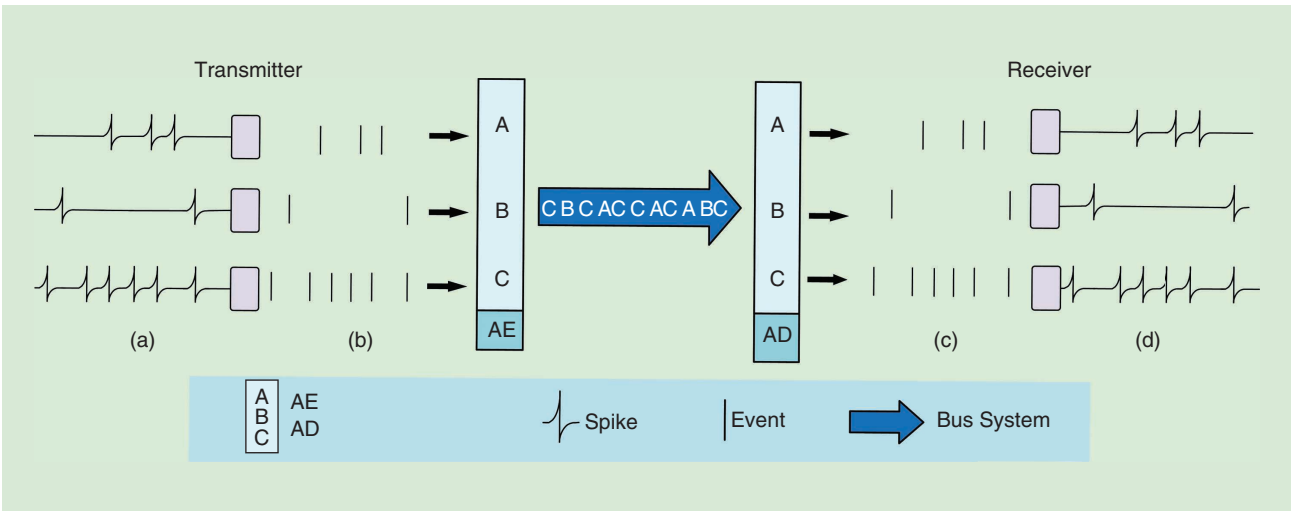

FIGURE 3. The AER communication protocol: (a) three neurons on the sending chip generate spikes; (b) spikes are interpreted as binary events. A binary address is generated by the AE and transmitted to the receiver chip by the bus line; (c) the binary address is decoded to the binary event by the AD; and (d) spikes are emitted on the corresponding neurons of the receiver chip where the positions of the neurons are determined by the AD.

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

where t_i is the time stamp of the event; t_n is the time stamp of the most recent neighborhood event; and d_t is the predefined threshold. The search for the most recent event checks eight neighborhood pixels around (x_i, y_i), as shown in Figure 4. It lacks temporal correlation with events in their spatial neighborhood because the event noise occurs randomly. Hence, the spatial-temporal correlation filter can effectively filter out event noise.

Motion consistency filter

In Figure 4, the principle of the motion consistency filter [13] is depicted. The blue dot denotes an incoming event caused by the object motion and the black dot represents an event noise. In the spatial-temporal domain, a newly incoming event should be consistent with the previous events (represented by red dots) caused by the same moving object. In a local region, the incoming event can be modeled as a consistent "moving plane" M. In this way, the velocity (v_x, v_y), can be used to assess the motion consistency, and the event noise can be removed because the previous events (the red dots, signal) and the black dot are not on the same plane. Concretely, the motion consistency plane for each active event e_i can be formulated as

$$ax_i + by_i + ct_i + d = 0, \tag{2}$$

where $(a, b, c, d) \in \mathbb{R}^4$ defines the plane M; (x_i, y_i) is the coordinate of event e_i; and t_i is the time stamp of event e_i. The event noise processing is an essential step to extract useful information from unwanted noise data for bio-inspired visual sensing and perception tasks of autonomous driving; it can promote the accuracy and speed of subsequent algorithms.

Event data representation

As an emerging sensing modality, event-based neuromorphic vision sensors only transmit local pixel-level changes caused by movement or light intensity change in a scene. The output data are sparse and asynchronous event streams which cannot be directly processed by standard vision pipelines, such as convolutional neural network (CNN)-based architecture. Therefore, encoding methods are utilized to convert asynchronous events into synchronous image- or grid-like representations for subsequent tasks such as object detection and tracking. According to whether or not the methods contain temporal information in the converted representations, we introduce two state-of-the-art encoding methods: spatial encoding and spatial-temporal encoding methods.

Spatial encoding

The spatial encoding methods convert event streams into event frames by storing event data at pixel location (x_i, y_i) with either fixed-time interval (e.g., 30 ms, constant time frame) or fixed number of events (e.g., 500 events, constant count frame). For an event frame, the value of the pixel is usually represented by the polarity of the last event (the positive event is 1 and the negative event is –1) or the statistical characteristics (such as the event count in a fixed-time interval, event count frame) of the events in the fixed interval. Assuming that $e_i(x_i, y_i, t_i, p_i)_{i\in[1,N]}$ represents event stream, typical approaches based on spatial encoding can be defined as follows:

1) *Constant time frames*:

$$F_j^t = \mathbf{card}(e_i | T \cdot (j-1) \le t_i \le T \cdot j), \tag{3}$$

where F_j^t represents the jth frame of time interval T; **card**() is the cardinality of a set; and e_i is the ith event of the event stream.

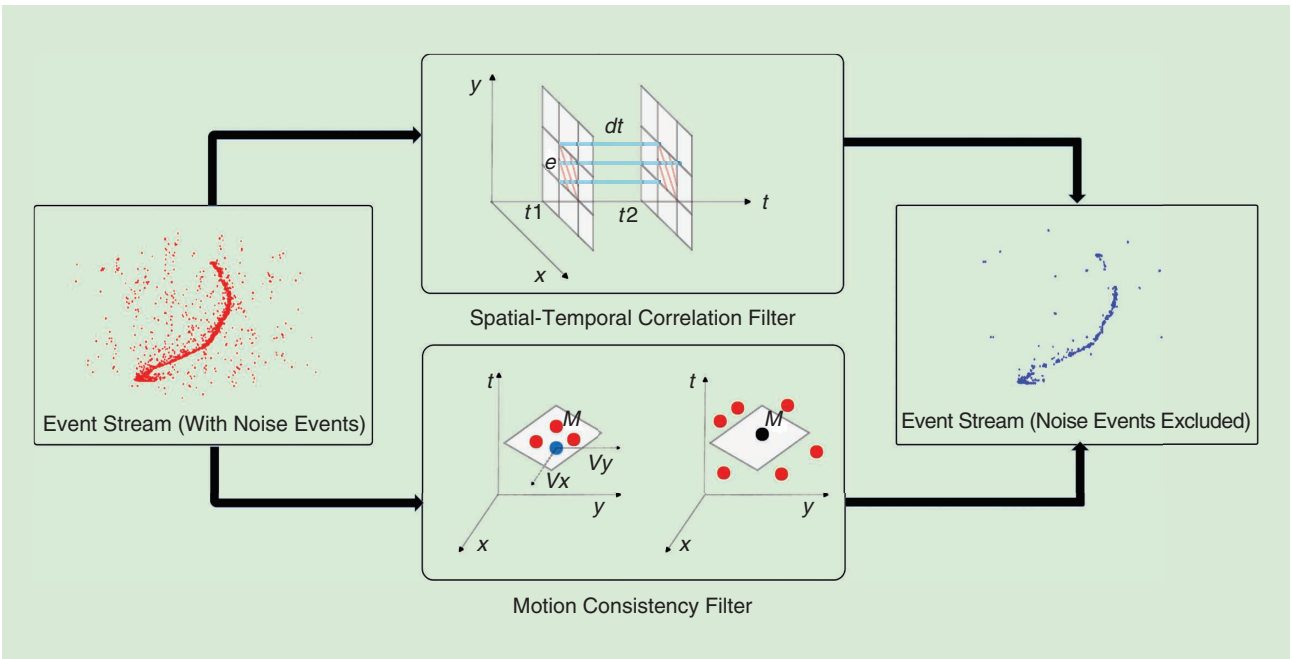

FIGURE 4. Event noise processing. The top branch is the spatial-temporal correlation filter; the bottom branch is the motion consistency filter.

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

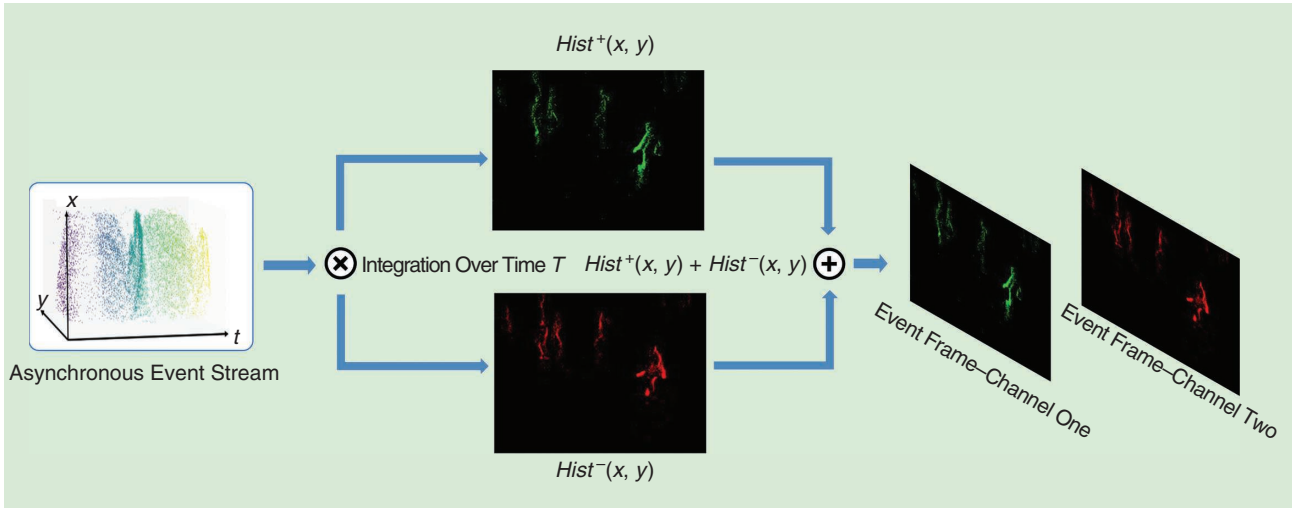

FIGURE 5. The process of converting asynchronous event data into an event frame. An event frame consists of two histograms from the positive events and negative events, respectively.

2) *Constant count frames*:

$$F_j^e = \mathbf{card}(e_i | E \cdot (j-1) \le i \le E \cdot j). \tag{4}$$

The constant count frame is defined similarly to constant time frame. F_j^e is the jth frame that contains E events.

3) *Event count frames*:

$$Hist^+(x, y) = \sum_{p_i=+1, t_i \in T} \delta(x - x_i, y - y_i). \tag{5}$$

Two separate histograms for positive and negative events are generated in a fixed-time interval T. $Hist^+(x, y)$ denotes the histogram for positive events, where δ is the Kronecker delta function. The same goes for the negative-events histogram, which is represented by $Hist^-$ with $p_i = -1$. The final representation of the events in the fixed-time interval T is an event frame, which consists of two histograms $Hist^+$ and $Hist^-$, as shown in Figure 5. Since the principle of the spatial encoding method is to project the events onto the spatial plane ($x - y$ plane), it loses the temporal information of all of the events.

Spatial-temporal encoding

The microsecond temporal resolution of the event stream provides a highly precise recording and description of the scene dynamics, which is valuable in many perception tasks such as high-speed moving object detection (e.g., vehicles). Spatial-temporal encoding methods combine spatial and temporal information of the events and convert events into a compact representation. A comparison of spatial-temporal encoding methods is presented in Table 1. A detailed description of these methods is displayed as follows.

Surface of active events

The surface of active events (SAE) uses time-stamp values instead of intensity values to represent the pixel values. For each incoming event e_i:

$$SAE: t_i \mapsto P(x_i, y_i), \tag{6}$$

where t_i is the time stamp of the most recent event at each pixel, the pixel value P at (x_i, y_i) is directly determined by the occurrence time of the events. The disadvantage of the SAE method is that it completely ignores the information of previous events happening at (x_i, y_i) and only uses the time stamp of the most recent event.

Leaky integrate-and-fire

Leaky integrate-and-fire (LIF) is an artificial neuron inspired by biological perception principles and computation primitives. A neuron receives input spikes (events) generated from a DVS, which modifies its membrane potential. If the membrane potential exceeds a predefined threshold, a spike stimulus will be sent to the output. The LIF neuron can be modeled as

Table 1. The comparison of different event data representations of spatial-temporal encoding.

Representation	Dimensions	Polarity Channel	Intensity	Weakness
SAE	H × W	2	Time stamp of the most recent event	Without temporal history
LIF	H × W	1	Event spikes	Without polarity information
Voxel grid	B × H × W	1	Sum event polarities	Without polarity information
EST	B × H × W	2	Sample event point-set into the grid	Without the least amount of information

The polarity channel is 2 if the encoding method considers the polarities of events; otherwise, it is 1. H and W represent the image height and width dimensions, respectively; B denotes the number of temporal bins.

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

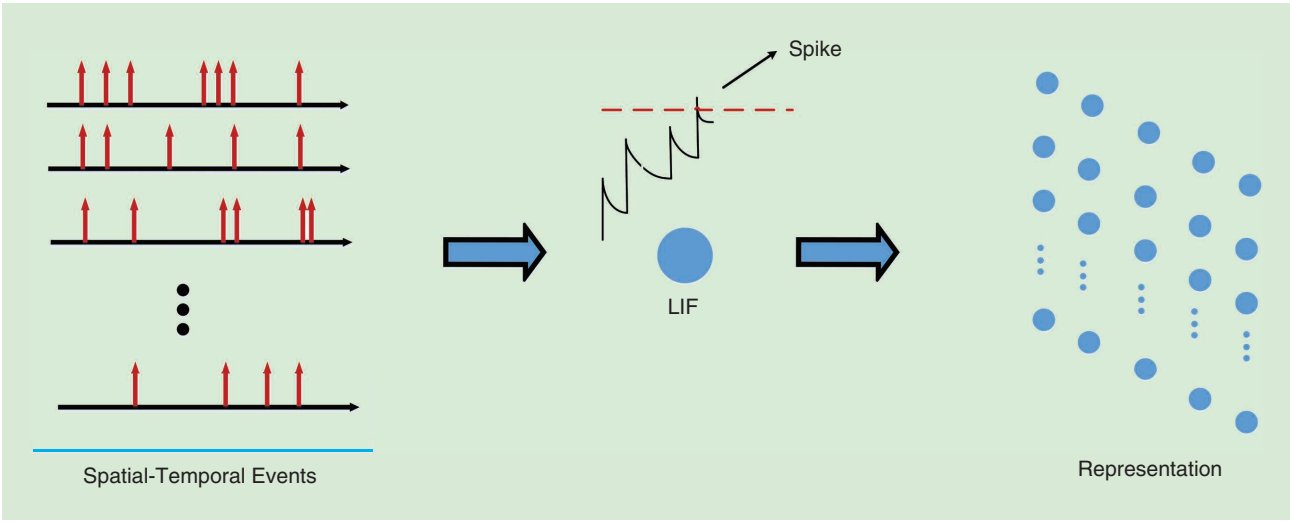

FIGURE 6. An LIF representation: Asynchronous spatial-temporal events are converted into event data representation by LIF neurons.

$$\tau \frac{dV}{dt} = -(V(t) - V_{\text{reset}}) + RI(t), \tag{7}$$

where, $V(t)$ is the membrane potential, which is a function across time; $I(t)$ is the total synaptic current; R is the membrane resistance; and τ is the membrane time constant. The neuron fires (produces an output spike) when the membrane potential reaches the threshold voltage (V_{th}) and then resets to reset voltage (V_{reset}). As shown in Figure 6, the spatial-temporal events are encoded by an LIF neuron, in which each event updates membrane potential of the neuron and the final converted representation is composed of the output spikes. An LIF neuron can not only transform event data into representation, it also serves as the basic unit of a spiking neural network (SNN) (see the section "SNNs").

Voxel grid

Voxel grid is a novel event representation aiming to improve the resolution of event stream in the temporal domain. Given a set of N events $(x_i, y_i, t_i, p_i)_{i \in [1,N]}$, B bins are used to split the time dimension; then, the time stamps of events are scaled to the range of $[0, B-1]$. The event voxel grid is defined as

$$\hat{t} = (B-1)(t_i - t_1)/(t_N - t_1), \tag{8}$$

$$V(x, y, t) = \sum_{i}^{N} p_i k(x - x_i) k(y - y_i) k(t - \hat{t}), \tag{9}$$

$$k(z) = \max(0, 1 - |z|), \tag{10}$$

where, $k(z)$ is the trilinear voting kernel, which is equivalent to the definition in [14]. As shown in Figure 7, events are converted into voxel grid representation with the fixed kernel. This representation retains the distribution of the events across the spatial-temporal dimensions.

Event spike tensor

Event spike tensor (EST) is an end-to-end learned representation [15]. In a given time interval T, EST can be formed by sampling the convolved signal,

$$S_{\pm}[x, y, t] = \sum_{e_i \in p_{\pm}} f_{\pm}(x_i, y_i, t_i) k_c(x - x_i, y - y_i, t - t_i), \tag{11}$$

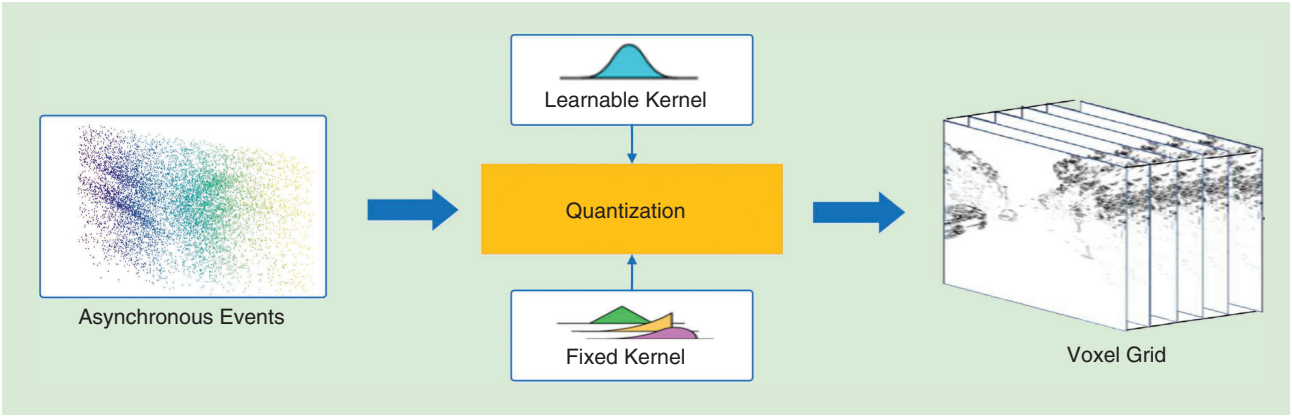

FIGURE 7. The process of converting asynchronous event data into grid-based representation with a fixed kernel [14] and a learnable kernel [15].

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

where, $f_{\pm}(x_i, y_i, t_i)$ is a measurement assigned to each event to represent the corresponding intensity value at the pixel location; k_c is the kernel convolution function to derive meaningful signal from the event stream. Generally, both the measurement and kernel are handcrafted functions in previous works, as illustrated in Figure 7. Particularly, the EST deploys a multilayer perception replacing the handcrafted kernel function in (11) to fit the data with the purpose of finding the best function for event streams. Simultaneously, the measurement function is chosen from a set of fixed functions. Examples of such function are the event polarity $f_{\pm} = \pm 1$; the event count $f_{\pm} = 1$; the time stamp $f_{\pm} = t$; and the normalized time stamp $f_{\pm} = (t - t_0)/T$.

There are many data sets that are built to promote the research of event-based neuromorphic vision, neurorobotics, and autonomous vehicles.

Event-based neuromorphic vision algorithms and applications of autonomous driving

The fundamental algorithms are the basis of the perception system of autonomous driving. For emerging systems and applications of bio-inspired vision, event-based neuromorphic vision algorithms are designed to extract features from event streams to fulfill given tasks. These methods can run directly on the event stream or take event representations as input (see the section "Event Data Representation"). They have been applied successfully in many vision tasks.

Event-based data sets of autonomous driving

In recent years, researchers have started to investigate the usage of event-based neuromorphic vision sensor such as DVS and DAVIS in the visual sensing and perception system of the autonomous driving system. There are many data sets that are built to promote the research of event-based neuromorphic vision, neurorobotics, and autonomous vehicles. In this section, four public event-based data sets dedicated to autonomous driving are discussed.

DET data sets

The performance of conventional lane extraction algorithms is limited because a frame-based camera cannot work well when the light is extremely dark or changes rapidly. To tackle this problem, [16] uses event-based neuromorphic vision sensors to build a high-resolution data set, called the *DET data sets*, for lane extraction. The DET data set containing various traffic scenes is collected by driving on tunnels, bridges, overpasses, and urban areas. The data set includes 5,424 event frames of 1,280 × 800 pixels with corresponding labels and consists of a training set of 2,716 frames, a validation set of 873 frames, and a test set of 1,835 frames. Two kinds of labels (per-pixel label without distinguishing lanes and per-pixel label with distinguishing lanes) are provided. The DET data set is the first bio-inspired vision data set for lane detection—a fundamental problem in autonomous driving.

N-CARS data sets

The N-CARS data set introduced by [17] provides recording cars in urban environments with a DVS. The data set consists of 12,336 car samples and 11,693 noncar (background) samples. Specifically, 7,940 car samples and 7,842 background samples are training samples, and others are testing samples. Each example is labeled by semiautomatic protocol with manual correction of the wrong one.

MVSEC data sets

In [4], the multivehicle stereo event camera data set (MVSEC) created for 3D perception with multiple sensors was presented. The MVSEC is the first data set with a synchronized stereo event-based neuromorphic vision system. The ground-truth depth data are generated from a calibrated lidar system contributing to stereo depth estimation with the event-based vision sensor. The MVSEC data set consists of long outdoor sequences in a variety of illuminations and driving speeds, which can be used for the evaluation of event-based visual odometry, localization, obstacle avoidance, and 3D reconstruction in challenging and real-word driving scenes.

DDD17 data set

For self-driving applications, end-to-end learning of the control model is a fascinating direction. The DDD17 data set [18] is the first large-scale public data set with a DAVIS sensor. The data are recorded in highway and city scenes driving from Switzerland to Germany. It has more than 12 h of data collected under different weather, road, and light conditions, covering the distance of more than 1,000 km. Furthermore, vehicle data, such as speed, GPS position, driver steering, throttle, and brake are also recorded.

Handcrafted feature

The concept of time surface is proposed to track the activity of the object due to the lack of effective low-level feature representations and descriptors for an event-based vision mission. It represents temporal characteristics and describes the spatial-temporal context around an event. For an event $e_i = (x_i, y_i, t_i, p_i)$, the time surface S_i of dimension $2R \times 2R$ is defined as

$$S_i = \begin{cases} e^{-\frac{t_i - T(C_i + R, P)}{\tau}}, & \text{if } p_i = P; \\ 0, & \text{otherwise,} \end{cases}$$

where $C_i = (x_i, y_i)$ is the pixel coordinates of the incoming event e_i, R is the radius of the spatial neighborhood around e_i, $T(C_i + R, P)$ is the time stamp of the last event with polarity P received from pixel $C_i + R$, and τ is a constant decay factor. The exponential decay expands the activity of past events and records history information of the activity in the neighborhood. Time surface has been effectively used in various vision tasks, such as object recognition and feature tracking. Further, a hierarchy of time surface is introduced for object recognition [19]. Relying on a time-oriented approach, this model is used to extract valuable spatial-temporal features from event

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

streams. Based on the findings in [19], a sparse coding basis decomposition was used to reduce the number of prototypes in a hierarchy structure for lowering computational cost and memory need [20]. However, they only achieved better recognition performance for simple shapes, such as numbers and letters, while they cannot well perform for complex objects, such as cars. Inspired by the histogram of orientation gradient feature widely used in frame-based vision, an effective event descriptor named *histogram of averaged time surfaces* (*HATS*) was constructed [17]. Then, better classification performance and real-time computation were obtained. HATS convert event streams into local memory time surfaces and computes the histograms to formulate the final descriptor. After these features are extracted from event streams, a simple linear support vector machine classifier is used to recognize objects in the N-CARS data set.

Clustering

A classical unsupervised learning approach is clustering. Given a set of data, the clustering algorithm can be used in this study to generate different groups. The data with different characteristics are grouped into different clusters. The clustering methods can be applied directly to generate object proposals because the event stream from the DVS can be treated as sparse point cloud data where each point is an event. For example, a clustering method named *Gaussian mixture models* (*GMMs*) is used to track the pedestrian [22]. The method achieves accurate detection and tracking of pedestrian objects by extending GMMs with a stochastic prediction of objects' states. The goal of tracking is to estimate the state of one or multiple objects over time. In case of a possible collision with other traffic participants, the autonomous vehicle requires sufficient reaction time to ensure a safe brake distance. It is difficult to track a pedestrian because a pedestrian can suddenly change his or her moving direction. The results in [22] indicated that applying clustering to spatial-temporal event data has a large potential for robust object tracking.

Bio-inspired feature learning

SNNs

An SNN is a bio-inspired approach that can operate directly on spatial-temporal event data. The computational pattern of SNNs mimics the working principle of receptive fields in the primary visual cortex. As basic building blocks of SNNs, LIF and adaptive exponential are both inspired by the biological neurons found in the visual cortex of mammalians, which encode temporal information and make them naturally fit asynchronous event streams. The basic principle of SNN is that a neuron will not emit any spike if it has not received any input spike from the preceding SNN layer. Moreover, the corresponding neuron will generate spikes that are fed to the next layer only if the membrane voltage caused by received spikes exceeds a predefined threshold. The predefined network units, such as the difference of Gaussians or Gabor filters, are usually used in the first layer of SNN to extract features. Features are transmitted from the first layer of SNN to the deeper layers in parallel [23]–[25]. The major disadvantage of conventional SNNs is not differentiable, causing the popular training methods to be inapplicable. In the context of autonomous driving, a SNN architecture consisting of refractory, convolution, and clustering layers was presented [26]. It was designed with biorealistic LIF neurons and synapses. The LIF neurons are used as basic building blocks in the proposed algorithm, where the refractory layer filters off fraction of the input events to generate spike. Then, the spikes are convolved by convolution layer to produce region proposal boxes. Moreover, the clustering layer combines these boxes to cluster together to form the shapes of objects. This method is validated on object detection with real traffic scenes including humans, bikes, cars, vans, trucks, and buses.

SNN with backpropagation

SNN with handcrafted feature extractors (such as Gabor filters) cannot learn weights naturally from the data. To overcome this drawback, researchers established a novel architecture of SNN with LIF neuron and winner-takes-all (WTA) circuits [21]. The LIF neuron uses dynamic weights rather than a simpler refractory mechanism to update its membrane potential. In a WTA circuit, it would inhibit other neurons from spiking once an output spike occurs in a neuron. Furthermore, the lateral inhibition is employed to put the dynamic weights of all inhibited neurons in the WTA circuit into the refractory state. The differentiable transfer functions are derived in the WTA configuration to make SNN trainable with backpropagation; moreover, the performance of SNN architecture is also improved. In Figure 8, an SNN network with backpropagation is illustrated. However, trainable SNN is only tested on simple data sets (such as MNIST) and has not been applied in specific autonomous driving scenarios. As the output of event-based neuromorphic vision sensor is a spatial temporal event stream which is fundamentally different from frame-based camera, it requires the design of specifically tailored algorithms to accommodate the nature of events, and [21] indicates the prospect of implementing deep SNNs.

CNN

CNN is a popular feature extraction architecture, which is composed of three types of layers, including a convolutional layer, a pooling layer, and a fully connected layer. It uses spatially localized convolutional filtering to capture local features of input image. Basic visual features, such as lines, edges, and corners, are learned in the first few layers, while more abstract features are learned in deeper layers. For an input image matrix I, the correspondence activation map M is computed in the nth neuron of the CNN as follows

$$M[i, j] = \sigma\left(\sum_{x=-2k-1}^{2k+1} \sum_{y=-2k-1}^{2k+1} W[x, y] I[i - x, j - y] + b \right), \quad (12)$$

where the image size is $2k + 1$, W is the nth convolutional filter, and σ is the nonlinear activation function. Generally, a max pooling layer follows each convolutional layer, in which

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

the local maximum is used to reduce the dimension of the matrix and prevent overfitting. Moreover, fully connected layers are usually added to learn the nonlinear combination of extracted features from previous layers. Over the decades, many variants of CNNs, such as fully CNNs and encoder–decoder networks, have emerged. These networks have different structures from traditional CNNs, such as removing the full connection layer. The performance of CNNs has surpassed traditional machine learning methods in many vision tasks, relying on successful training algorithms and large amounts of data.

CNNs for optical flow, depth, and egomotion

Known as a 2D motion estimation, the optical flow is defined as the distribution of apparent velocities of movement of brightness patterns between two images. It provides valuable information about the scene and serves as input for several tasks, such as tracking and visual odometry. In the neuromorphic vision research community, some works attempt to estimate optical flow by taking advantage of high temporal resolution of event-based sensors [27]. EV-FlowNet, a self-supervised deep learning architecture for optical flow estimation for event-based sensors, is proposed in [28]. In this method, a four-channel event representation consisting of the histogram (5) and SAE (6) of different polarity is used to pass through a pipeline that is composed of four stride convolutional layers, two residual blocks, and four up-sampling convolutional layers for obtaining flow estimation. By evaluating an MVSEC data set, the network is able to accurately predict optical flow from event streams. In [29], a novel neural network framework is proposed to acquire motion information including optical flow, depth, and egomotion from a set of inputs (a voxel grid) that is an event data representation mentioned in the "Voxel Grid" section. The network architecture consists of encoder–decoder networks and pose models; among them, the encoder–decoder section is responsible for predicting optical flow and depth, while the pose model is responsible for estimating egomotion. Experimental results in the MVSEC data set indicate that the presented network can learn various motion information of events well. Recently, a lightweight evenly cascaded convolutional network (ECN) using monocular event-based sensor input for dense depth, optical flow, and egomotion estimation was introduced in [30]. ECNs use an encoder network to predict pose; meanwhile, an encoder–decoder network is applied to obtain the scaled depth. The algorithm can operate at 250 frames/s (fps) on a single NVIDIA 1,080 titanium GPU. Compared with previous works, it makes significant improvements on the performance of the MVSEC data set.

CNNs for object detection

Reliable object detection is essential to avoid accidents that might be life threatening because a self-driving car is sharing the road with many traffic participants, such as vehicles and pedestrians. For instance, a supervised learning method is applied on event data for object detection under egomotion [31]. The data set used in this article is DDD17, which is a large event-based data set applying DAVIS to record various challenging scenarios under egomotion. The DAVIS is a sensor consisting of an event-based neuromorphic sensor and a synchronized gray-scale frame-based camera. In [31], gray-scale images are fed into a state-of-the-art frame-based CNN to generate outputs (pseudolabels), which are used as ground truths for subsequent training on event-based data. This method achieves high-speed detection (100 fps) in a real outdoor scenario within various backgrounds such as day and night. As pseudolabels are not explicit enough, the authors manually labeled the DDD17 data set to explore the potential of event-based neuromorphic sensor for vehicle detection in autonomous driving [32]. A convolutional SNN is utilized to generate visual attention maps for synchronizing with the frame-based stream. Two separate event-based and frame-based streams are incorporated into a CNN detector to obtain detection output. With a joint decision model to postprocess the output, the algorithm outperforms the state-of-the-art

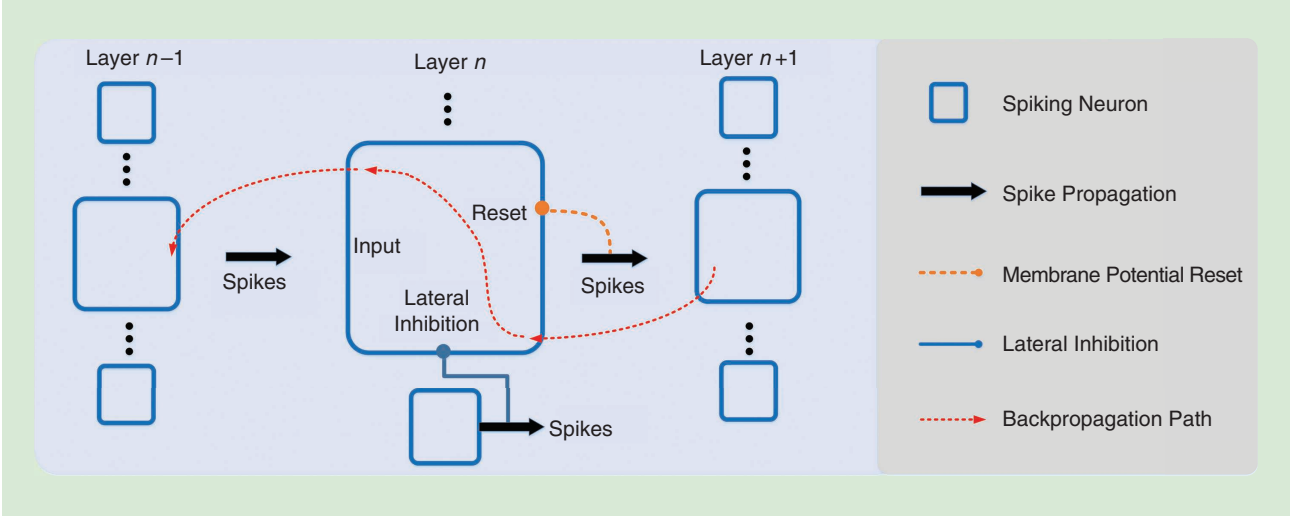

FIGURE 8. An example of how an SNN network works with backpropagation [21].

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

methods that only employ frame-based cameras. The detection for stationary and moving people around a self-driving car has attracted the attention of researchers. Specifically, a multicue event information fusion for pedestrian detection was proposed [33]; it was evaluated on the data set recorded by a neuromorphic vision sensor. Based on the advantages of leveraging various properties of event streams, this article performed better on positioning and recognition of pedestrians. Recently, a cross-modal approach was presented in [34]; wormhole learning was utilized to pair red, green, blue (RGB) camera and event-based neuromorphic vision sensors to improve the object detection performance under the scenario of urban driving. This method is different from transfer learning as it can be transferred back to the original domain to improve performance on the task. The experimental results of wormhole learning reveal that there are many innovative approaches to combine data from different heterogeneous sensors, such as RGB cameras, infrared cameras and neuromorphic vision sensors.

CNNs for semantic segmentation

In the sensing and perception system of autonomous driving, a comprehensive understanding of the surrounding environment is provided by semantic segmentation. The first CNN-based baseline for semantic segmentation with an event-based neuromorphic vision sensor is introduced in [35]. In this article, the authors build an event-segmentation data set (Ev-Seg) that is an extended version of DDD17 for semantic segmentation. Inspired by the study in [31], the labels of Ev-Seg are generated by running a trained CNN on gray-scale images. Then, an Xception-based CNN architecture is trained to learn generalization ability from event streams. Finally, the complementarity between the frame-based camera and event-based neuromorphic vision sensor is presented through comparing the semantic segmentation results produced from event data and corresponding gray-scale images.

CNNs for active perception

Controlling the autonomous vehicle in challenging scenes such as highway driving requires a low-latency perception system. Hence, researchers try to tackle this tough problem by unlocking the low-latency potential of event-based neuromorphic vision sensors. An end-to-end autonomous driving system, mapping from the event streams to the driving actions, is proposed in [3]. This system converts events to event count frames (histogram of different polarities) mentioned in (5), which are fed into a residual neural network (ResNet)-inspired network to predict the steering angle of the vehicle. The proposed method can accurately predict the steering angle of vehicles and performs better on DDD17 data sets than the state-of-the-art systems using gray-scale images.

CNNs-to-SNNs

CNNs have demonstrated their ability to deal with many difficult vision problems, such as object detection. SNNs have presented their potential for low-power event-driven neuromorphic hardware. However, the applications of SNNs are limited due to their shallow neural network architecture. Furthermore, the CNN-to-SNN model is developed to combine the benefits of deep architecture in CNNs with the bio-inspired mechanism of SNNs. References [36]–[38] illustrate that widely used CNNs, such as VGG, ResNet, and Inception-V3 can be converted into spiking networks. It is worth mentioning that the network can achieve a more robust performance via conversion from CNNs, although the conversion process would lose some precision and increase computation. Some works have been reviewed in [39].

Transfer learning via pretrained network

Transfer learning is a very effective method to improve the training performance of the deep neural network. Knowledge learned from a different domain can be exploited to initialize the weights of a deep neural network. The availability of event-based data sets collected with a DVS sensor is limited compared with the data set recorded by frame-based cameras. Thus, by starting the supervised training process from a better set of initial weights, the requirement of the training data can be reduced, and the generalization ability of the network can be improved. Pretrained models, such as VGGnet and ResNet, can be applied to bio-inspired sensing and perception tasks of autonomous driving. Specifically, event streams can be transformed into a three-channel image-like representation to serve as input to pretrained CNNs. In [40], the authors combined an inceptive event time surface (IETS) with transfer learning to improve performance of object classification. IETSs are generated to utilized transfer learning from the GoogLeNet that is pretrained on ImageNet, including the millions of real-world images. Nearly 100% classification accuracy on the event-based N-CARS data set is achieved by the algorithm. In [41], a robust event stream object tracking method is presented. A VGG-16 model pretrained on ImageNet is used to extract features to represent the appearance of the object. Based on correlative filter mechanism, the correlation response map is computed on the extracted features. The proposed approach performs well in various challenging visual scenarios.

Event-based assistance systems

After the basics of event-based perception system of autonomous driving are covered, the event-based assistance systems are discussed.

Image reconstruction

The event-based neuromorphic vision sensor generates HDR event data even in extreme illumination conditions and also avoids motion blur under rapid motion. Reconstructing HDR intensity images from event streams facilitates the adoption of mature computer vision techniques. Previous works focus on exploiting the low latency of neuromorphic vision sensor by directly processing event data (such as SNNs) or transferring events to image-like or grid-like representations as mentioned in the section "Event Data Representation." However, the deep

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

neural network trained on real image data (such as ImageNet) cannot be effectively transferred to these representations, even though it achieves some performance improvements (see the section "Transfer Learning via Pretrained Network"). As an alternative method, image reconstruction (IR) from event streams is first proposed in [42]. IR can achieve both high frame-rate images and high-quality images with no motion blur. In [43], the authors utilize the time stamp of new events to define a manifold for IR. With considering IR as an energy minimization problem, the proposed method is optimized and achieves real-time performance on a GPU. Furthermore, an asynchronous complementary filter is presented to reconstruct event streams for continuous-time intensity estimate [44]. In this article, the gray-scale frames and events produced by DAVIS are fused into an image with high temporal resolution and HDR. In addition, a new framework for IR, named *E2VID*, is introduced in [45]. E2VID converts event stream into 3D spatial-temporal voxel grid sequences (see the "Voxel Grid" section), which are taken as the input of the network. The algorithm is trained on a large synthetic event data simulated with ESIM [46] to generate reconstructed image frames. The reconstructed image data from event streams can be used for various applications such as object recognition, SLAM, and optical flow estimation.

Event-based neuromorphic vision is an emerging technique in the era of mature sensor hardware of autonomous driving.

Panoramic stereo vision

Panoramic vision in 3D offers a full 360° surrounding view which facilitates the navigation and localization tasks for autonomous driving. A novel multiperspective panoramic stereo event-based vision system is proposed in [47]. It is composed of a pair of line event-based neuromorphic vision sensors. The authors present a novel event-driven stereo matching approach for 3D panoramic vision. The process steps of the event-driven stereo matchintg algorithm include event map generation, event distribution measure, cost calculation, disparity estimation and refinement. The experimental results indicate that the tailored event-driven stereo method achieves accurately 3D reconstruction in real time out of 360° panoramic views.

Visual odometry

The goal of the visual odometry is to estimate the position and orientation of a vehicle with vision sensors. The visual odometry system of an autonomous vehicle with a traditional frame-based camera has been developed for many years, while the method based on an event-based neuromorphic vision sensor is still in the preliminary stage. For example, an event-based visual odometry system for intelligent vehicle applications is proposed in [48]. The events generated from a DAVIS sensor are aggregated into constant time frame defined in (3) to serve as input to subsequent algorithms. The feature tracking is used by visual odometry system to develop parallel pose estimation and mapping. The feasibility of event-based neuromorphic vision sensors for bio-inspired visual odometry systems in real-word outdoor driving scenes is confirmed by the results of their experiment on the MVSEC data set.

Drowsiness driving monitoring

Drowsiness driving monitoring is important to ensure that the autonomous driving vehicle is under the supervision of the drivers. In [49], an event-based drowsiness driving detection system is proposed. The event-based neuromorphic vision sensor is considered as an efficient and effective detector for the drowsiness driving-related motions due to the unique output. [49] proposes to recognize and localize the driver's eyes and mouth motions from event streams, and extracts event-based drowsiness-related features directly from the event streams caused by eye and mouth motions. Experiments in [49] demonstrate the high efficiency and accuracy under different illumination conditions such as subjects wearing sunglasses.

Spike compression

The event data compression is particularly important for maintaining the real-time performance of the sensing system of autonomous vehicles because both the data storage and transmission bandwidth of on-board event-based neuromorphic visions sensors equipped on the autonomous vehicles are limited. To address this problem, a cube-based spike coding framework is proposed by [50]. In the spatial-temporal dimension, an octree-based structure is put forward to adaptively cut the event (spike) stream into coding cubes, then address-prior mode and time-prior mode are designed to exploit the spatial and temporal characteristics of events for data compression. The proposed spike coding framework is evaluated on the DDD17 data set. Experimental results indicate that it can achieve a better compression ratio against the raw event data. Reference [51] proposes to use mixture density autoencoder to learn a low-dimensional representation from an event stream, which preserves the nature of event-based data better while being easy to feed to a sequence classifier.

Challenges and future directions in autonomous driving

Event-based neuromorphic vision is an emerging technique in the era of mature sensor hardware of autonomous driving. Comparing it with lidar, radar, and cameras is unfair because event-based sensors such as DVS are not at the same maturity level as others. Conversely, there is substantial room for the development and improvement in the cross-research of event-based neuromorphic vision and autonomous driving. Challenges and future directions closely related to autonomous driving are pointed out in numerous opportunities, as described later.

Sensor fusion in perception system of autonomous driving

To fuse the event-based neuromorphic vision sensor with others, there is an unavoidable problem that the sensor fusion brings back the disadvantages of providing a redundant,

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

sampled intensity output with linear encoding of intensity. On the contrary, the advantages are also obvious; that is, different kinds of sensors are complementary. For example, DVS contains no color information, which is provided by frame-based cameras. The distance and speed information can be provided by lidar and radar. It remains to be seen whether the DVS output can be used to trigger frame captures of other sensors. If it is, the DVS and other sensors can operate together with mixed conventional machine vision, bio-inspired, and event-based neuromorphic vision-based approaches. Therefore, some of the limitations of a traditional sensor-based perception system may be overcome; moreover, new scenarios that were previously inaccessible in the visual sensing and perception of autonomous vehicles might be reached.

There is an emerging need for high-quality benchmarks in the fields of event-based neuromorphic vision and autonomous driving.

Active vision system of autonomous driving

In robotics, the ability to directly fuse the perception with its motoric ability is often referred to as *active perception*. In autonomous driving, it is found that the perception and action are often kept in separated spaces; this is a consequence of state-of-the-art sensors equipped on the autonomous vehicle being frame-based. The sensing and perception only exist in a discrete moment while the motion is a continuous entity. It can be argued that the event-based neuromorphic vision sensor can see the motion, which has the potential to cross the bridge between perception and motor control. New methods of encoding perceptions and actions could be meaningful to the active perception system of autonomous driving. Moreover, this would create new opportunities for real-time navigation and obstacle avoidance for autonomous driving if the visual perception can be bound with the system dynamic to enable dynamic environment perception.

Large-scale autonomous driving benchmark based on an event-based neuromorphic vision sensor

It is well known that rapid development of autonomous driving is promoted by standardized benchmarks. For example, the growing popularity of deep neural networks in intelligent vehicles and large-scale benchmarks such as KITTI, Cityscale, and ImageNet, is interconnected and mutually reinforced. In the earlier days of event-based neuromorphic vison, most of the research work was done in an indoor environment due to the low resolution of sensors. Until recently, the event-based neuromorphic vision sensor has been expanded to outdoor scenarios, such as autonomous driving, by the teams of Tobi Debruck, Kostas Daniilidis, and David Scaramuzzsa. There is an emerging need for high-quality benchmarks in the fields of event-based neuromorphic vision and autonomous driving. A standard platform would bring the mainstream of computer vision-based intelligent vehicle research to pay attention to event-based neuromorphic vision; furthermore, the unique strengths of bio-inspired vision would be leveraged to attract research interests in new sensing techniques for autonomous driving.

From simulated event data to real-world autonomous driving

Labeling the asynchronous event data is always a challenging problem because almost all of the annotation tools are developed for frame-based cameras. Additionally, there is not a standard format for the annotations. From one perspective, developing an easy-to-use tool for recording and labeling event data would make a significant contribution to the community; from another perspective, the adoption of event-based neuromorphic vision technology would also be facilitated by developing simulators. Particularly, the corresponding event streams, intensity frames, and depth information could be generated by a simulator based on the working principle of the sensor. Simultaneously, the basic facts of all recording data including the trajectory of the sensor, the label of the object, and even the optical flow are also generated without the need for annotation. With photo-realistic virtual driving scenes and realistic sensor models, the development of event-based visual sensing and perception system in autonomous vehicles will be accelerated by prototyping on simulated event data with transfer learning methods in the future.

Limitations that may exist as event-based neuromorphic vision sensors mature

There is no appearance feature such as color and texture because an event-based neuromorphic vision sensor only transmits local pixel-level changes, making it perform poorly in some applications with high requirements for appearance features. Although researchers have used the method of IR (mentioned in the section "Spatial Encoding") to reconstruct image frames from event streams, the quality of reconstructed image frames is still not comparable to the output data produced by RGB cameras. The application of an event-based neuromorphic vision sensor is limited in some scenarios where energy, latency, and dynamic range are not important, especially in high-resolution complex scenarios.

Conclusions

Innovative solutions will emerge due to the challenges remaining on the road to fully autonomous driving. Concurrently, sophisticated signal processing techniques have been successfully applied to autonomous driving hardware such as cameras, lidars, and radars. Exploring alternative methods of visual sensing such as event-based neuromorphic vision is promising for promoting subsequent tasks to be more robust and complementary. It is reasonable to say that the research and development of an event-based neuromorphic vision for autonomous driving is still in its infancy. In this article, the advantages, signal processing techniques, emerging applications and systems, and future directions of an event-based neuromorphic vision for autonomous driving have been introduced and analyzed. This article helps researchers and engineers take the first step in developing innovative signal

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

processing techniques toward bio-inspired visual sensing and perception of autonomous vehicles.

Acknowledgments

This research has received funding from the Shanghai Automotive Industry Sci-Tech Development Program according to grant agreement 1838, from the Shanghai AI Innovation Development Program 2018, and from the European Union's Horizon 2020 Framework Program for Research and Innovation under the specific grant agreement 785907 (Human Brain Project SGA2).

Authors

Guang Chen (guangchen@tongji.edu.cn) received his B.S and M.Eng. degrees from Hunan University and his Ph.D. degree from the Technical University of Munich (TUM). He is a research professor at Tongji University, where he leads the Intelligent Perception and Intelligent Computation group. He is also a senior research associate (guest) at TUM. He was a research scientist at fortiss GmbH from 2012 to 2016, and a senior researcher at the Chair of Robotics, Artificial Intelligence and Real-Time Systems at TUM from 2016 to 2017. He was named the Tongji Hundred Talent Research Professor 2018. His research interests include computer vision, machine learning, and bio-inspired vision with applications in robotics and autonomous vehicles. He is a Member of the IEEE.

Hu Cao (hu.cao@tum.de) received his B.S. degree in vehicle engineering from Anhui University of Technology, China, in 2017 and his M.Eng. degree in vehicle engineering from Hunan University in 2019. He is currently a Ph.D. candidate at the Technical University of Munich. His research interests include neuromorphic engineering, robotics, and deep learning.

Jörg Conradt (conr@kth.se) received his M.S. degree in robotics from the University of Southern California and a diploma in computer engineering from the Technical University of Berlin. He received his Ph.D. degree from ETH Zurich. He an associate professor at the KTH School of Electrical Engineering and Computer Science in Stockholm, Sweden. Before joining KTH, he was a W1 professor at the Technical University of Munich (TUM). He was the founding director of the Elite Master Program Neuroengineering at TUM. He is a Senior Member of the IEEE.

Huajin Tang (htang@zju.edu.cn) received his B.Eng. degree from Zhejiang University, his M.Eng. degree from Shanghai Jiao Tong University, and his Ph.D. degree from the National University of Singapore. Since 2008, he has been the head of the Robotic Cognition Lab, Institute for Infocomm Research, A*STAR, Singapore. Since 2014, he has been a professor at Sichuan University and is now a professor at Zhejiang University, China. He received the 2016 IEEE Transactions on Neural Networks and Learning Systems Outstanding Article Award and 2019 IEEE Computational Intelligence Magazine Outstanding Article Award. His research interests include neuromorphic computing, neuromorphic hardware, and robotic cognition, among others. He has served as an associate editor of *IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cognitive and Developmental Systems, Frontiers in Neuromorphic Engineering*, and *Neural Networks*. He is a Board of Governors member of the International Neural Networks Society. He is a Member of the IEEE.

Florian Röhrbein (florian@gmx.org) received his diploma and Ph.D. degree from the Technical University of Munich (TUM) and the venia legendi for computer science from the University of Bremen. He is responsible for the development and implementation of the artificial intelligence (AI) strategy for a world-leading company and is the chief editor of *Frontiers in Neurorobotics*. He was also the managing director in the Human Brain Project at TUM. He has international experience in various projects on AI, computational neuroscience and brain-inspired cognitive systems. Research stays include the MacKay Institute of Communication and Neuroscience (United Kingdom), the Honda Research Institute Europe, and the Albert Einstein College of Medicine (New York). He is a Senior Member of the IEEE.

Alois Knoll (knoll@in.tum.de) received his M.S. degree in electrical/communications engineering from the University of Stuttgart, Germany, in 1985 and his Ph.D. degree (summa cum laude) in computer science from the Technical University (TU) of Berlin, Germany, in 1988. He served on the faculty of the Department of Computer Science of TU Berlin until 1993. He joined the University of Bielefeld as a full professor and was the director of the research group Technical Informatics until 2001. Since 2001, he has been a professor at the Department of Informatics, Technical University of Munich (TUM), and was executive director of the Institute of Computer Science at TUM from 2004 to 2006. He was the program chair of IEEE Humanoids2000, general chair of IEEE Humanoids2003, program chair of IEEE-IROS 2015, and editor-in-chief of *Frontiers in Neurorobotics*. He is a Senior Member of the IEEE.

References

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008. doi: 10.1109/JSSC.2007.914337.

[2] S. Liu, B. Rueckauer, E. Ceolini, A. Huber, and T. Delbruck, "Event-driven sensing for efficient perception: Vision and audition algorithms," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 29–37, Nov. 2019. doi: 10.1109/MSP.2019.2928127.

[3] 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 Conf. Computer Vision and Pattern Recognition*, 2018, pp. 5419–5427. doi: 10.1109/CVPR.2018.00568.

[4] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2032–2039, 2018. doi: 10.1109/LRA.2018.2800793.

[5] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. N. Clark, J. Dolan, D. Duggins et al., "Autonomous driving in urban environments: Boss and the urban challenge," *J. Field Robot.*, vol. 25, no. 8, pp. 425–466, 2008. doi: 10.1002/rob.20255.

[6] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. F. R. Jesus, R. F. Berriel et al., Self-driving cars: A survey. 2019. [Online]. Available: arXiv:1901.04407

[7] E. Guizzo, "How Google's self-driving car works," IEEE Spectrum, Oct. 18, 2011. [Online]. Available: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/how-google-self-driving-car-works

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.

[8] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison et al., Event-based vision: A survey. 2019. [Online]. Available: arXiv:1904.08405

[9] M. A. Mahowald and C. Mead, "The silicon retina," *Sci. Amer.*, vol. 264, no. 5, pp. 76–82, 1991. doi: 10.1038/scientificamerican0591-76.

[10] H. Liu, C. Brandli, C. Li, S. Liu, and T. Delbruck, "Design of a spatiotemporal correlation filter for event-based sensors," in *Proc. 2015 IEEE Int. Symp. Circuits and Systems (ISCAS)*, pp. 722–725. doi: 10.1109/ISCAS.2015.7168735.

[11] V. Padala, A. Basu, and G. Orchard, "A noise filtering algorithm for event-based asynchronous change detection image sensors on TrueNorth and its implementation on TrueNorth," *Front. Neurosci.*, vol. 12, p. 118, 2018. doi: 10.3389/fnins.2018.00118.

[12] A. Khodamoradi and R. Kastner, "O(n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors," *IEEE Trans. Emerg. Topics Comput.*, to be published. doi: 10.1109/TETC.2017.2788865.

[13] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen, "EV-gait: Event-based robust gait recognition using dynamic vision sensors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 6351–6360. doi: 10.1109/CVPR.2019.00652.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Advances Neural Information Processing Systems 28*, 2015, pp. 2017–2025.

[15] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, 2019, pp. 5632–5642, doi: 10.1109/ICCV.2019.00573.

[16] W. Cheng, H. Luo, W. Yang, L. Yu, S. Chen, and W. Li, "DET: A high-resolution DVS dataset for lane extraction," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019, pp. 1–10.

[17] 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. 2018 IEEE/Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 1731–1740. doi: 10.1109/CVPR.2018.00186.

[18] J. Binas, D. Neil, S. Liu, and T. Delbrück, "DDD17: End-to-end DAVIS driving dataset," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, 2017, pp. 1–9.

[19] 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, July 2017. doi: 10.1109/TPAMI.2016.2574707.

[20] G. Haessig and R. Benosman, "A sparse coding multi-scale precise-timing machine learning algorithm for neuromorphic event-based sensors," in *Proc. Micro- and Nanotechnology Sensors, Systems, and Applications X*, 2018, vol. 10639, pp. 289–296. doi: 10.1117/12.2305933.

[21] J. H. Lee, T. Delbruck, and M. Pfeiffer, "Training deep spiking neural networks using backpropagation," *Front. Neurosci.*, vol. 10, p. 508, 2016. doi: 10.3389/fnins.2016.00508.

[22] E. Piatkowska, A. N. Belbachir, S. Schraml, and M. Gelautz, "Spatiotemporal multiple persons tracking using dynamic vision sensor," in *Proc. 2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops*, pp. 35–40. doi: 10.1109/CVPRW.2012.6238892.

[23] G. Orchard, C. Meyer, R. Etienne-Cummings, C. Posch, N. Thakor, and R. Benosman, "HFirst: A temporal approach to object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2028–2040, Jan. 2015. doi: 10.1109/TPAMI.2015.2392947.

[24] F. Folowosele, R. J. Vogelstein, and R. Etienne-Cummings, "Towards a cortical prosthesis: Implementing a spike-based HMAX model of visual object recognition in silico," *IEEE J. Emerg. Select. Topics Circuits Syst.*, vol. 1, no. 4, pp. 516–525, Dec. 2011. doi: 10.1109/JETCAS.2012.2183409.

[25] 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.*, to be published. doi: 10.1109/TNNLS.2019.2945630.

[26] J. Acharya, V. Padala, and A. Basu, "Spiking neural network based region proposal networks for neuromorphic vision sensors," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, May 2019, pp. 1–5. doi: 10.1109/ISCAS.2019.8702651.

[27] R. Benosman, C. Clercq, X. Lagorce, S. Ieng, and C. Bartolozzi, "Event-based visual flow," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 407–417, Feb. 2014. doi: 10.1109/TNNLS.2013.2273537.

[28] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "EV-flownet: Self-supervised optical flow estimation for event-based cameras," in *Proc. Robotics: Science and System*, 2018, pp. 1–9. doi: 10.15607/RSS.2018.XIV.062.

[29] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 989–997. doi: 10.1109/CVPR.2019.00108.

[30] C. Ye, A. Mitrokhin, C. Parameshwara, C. Fermüller, J. A. Yorke, and Y. Aloimonos, Unsupervised learning of dense optical flow and depth from sparse event data. 2018. [Online]. Available: arXiv:abs/1809.08625

[31] N. F. Y. Chen, "Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion," in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 757–766. doi: 10.1109/CVPRW.2018.00107.

[32] J. Li, S. Dong, Z. Yu, Y. Tian, and T. Huang, "Event-based vision enhanced: A joint detection framework in autonomous driving," in *Proc. 2019 IEEE Int. Conf. Multimedia and Expo (ICME)*, pp. 1396–1401. doi: 10.1109/ICME.2019.00242.

[33] G. Chen, H. Cao, C. Ye, Z. Zhang, X. Liu, X. Mo, Z. Qu, J. Conradt et al., "Multi-cue event information fusion for pedestrian detection with neuromorphic vision sensors," *Front. Neurorobot.*, vol. 13, p. 10, Apr. 2019. doi: 10.3389/fnbot.2019.00010.

[34] A. Zanardi, A. Aumiller, J. Zilly, A. Censi, and E. Frazzoli, "Cross-modal learning filters for RGB-neuromorphic wormhole learning," in *Proc. 15th Robotics: Science and System XV*, June 24, 2019, p. P45.

[35] I. Alonso and A. C. Murillo, "EV-SegNet: Semantic segmentation for event-based cameras," *2019 IEEE/Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–10.

[36] Y. Cao, Y. Chen, and D. Khosla, "Spiking deep convolutional neural networks for energy-efficient object recognition," *Int. J. Comput. Vis.*, vol. 113, no. 1, pp. 1573–1405, 2015. doi: 10.1007/s11263-014-0788-3.

[37] B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu, "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification," *Front. Neurosci.*, vol. 11, p. 682, Dec. 2017. doi: 10.3389/fnins.2017.00682.

[38] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," *Front. Neurosci.*, vol. 13, p. 95, Mar. 2019. doi: 10.3389/fnins.2019.00095.

[39] M. Pfeiffer and T. Pfeil, "Deep learning with spiking neurons: Opportunities and challenges," *Front. Neurosci.*, vol. 12, p. 774, Oct. 2018. doi: 10.3389/fnins.2018.00774.

[40] R. Wes Baldwin, M. Almatrafi, J. R. Kaufman, V. Asari, and K. Hirakawa, "Inceptive event time-surfaces for object classification using neuromorphic cameras," in *Proc. Int. Conf. Image Analysis and Recognition*, 2019, pp. 395–403. doi: 07/978-3-030-27272-2_35.

[41] H. Li and L. Shi, "Robust event-based object tracking combining correlation filter and CNN representation," *Front. Neurorobot.*, vol. 13, p. 82, Oct. 2019. doi: 10.3389/fnbot.2019.00082.

[42] M. Cook, L. Gugelmann, F. Jug, C. Krautz, and A. Steger, "Interacting maps for fast visual interpretation," in *Proc. 2011 Int. Joint Conf. Neural Networks*, pp. 770–776. doi: 10.1109/IJCNN.2011.6033299.

[43] G. Munda, C. Reinbacher, and T. Pock, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1381–1393, 2018. doi: 10.1007/s11263-018-1106-2.

[44] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *Computer Vision—ACCV 2018*. New York: Springer-Verlag, 2019, pp. 308–324.

[45] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, 1–23.

[46] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. 2nd Conf. Robot Learning*, Oct. 29–31, 2018, vol. 87, pp. 969–982.

[47] S. Schraml, A. N. Belbachir, and H. Bischof, "Event-driven stereo matching for real-time 3D panoramic vision," in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 466–474. doi: 10.1109/CVPR.2015.7298644.

[48] D. Zhu, J. Dong, Z. Xu, C. Ye, Y. Hu, H. Su, Z. Liu, and G. Chen, "Neuromorphic visual odometry system for intelligent vehicle application with bio-inspired vision sensor," in *Proc. IEEE Int. Conf. Robotics and Biomimetics*, Sept. 2019, pp. 2225–2232. doi: 10.1109/ROBIO49542.2019.8961878.

[49] 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.*, to be published. doi: 10.1109/JSEN.2020.2973049.

[50] S. Dong, Z. Bi, Y. Tian, and T. Huang, "Spike coding for dynamic vision sensor in intelligent driving," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 60–71, Feb. 2019. doi: 10.1109/JIOT.2018.2872984.

[51] G. Chen, J. Chen, M. Lienen, J. Conradt, F. Röhrbein, and A. C. Knoll, "FLGR: Fixed Length Gists Representation learning for RNN-HMM hybrid-based neuromorphic continuous gesture recognition," *Front. Neurosci.*, vol. 13, p. 73, Feb. 2019. doi: 10.3389/fnins.2019.00073.

[52] A. Zanardi, J. Zilly, A. Aumiller, A. Censi, and E. Frazzoli, "Wormhole learning," in *Proc. 2019 Int. Conf. Robotics and Automation (ICRA)*, Montreal, May 20–24, 2019, pp. 7899–7905. doi: 10.1109/ICRA.2019.8794336.

[53] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck, "Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output," *Proc. IEEE*, vol. 102, no. 10, pp. 1470–1484, Oct. 2014. doi: 10.1109/JPROC.2014.2346153.

SP

Authorized licensed use limited to: Technische Universitaet Muenchen. Downloaded on July 27,2022 at 10:04:39 UTC from IEEE Xplore. Restrictions apply.