NeuroIV: Neuromorphic Vision Meets Intelligent Vehicle Towards Safe Driving With a New Database and Baseline Evaluations

Guang Chen, *Member, IEEE*, Fa Wang, Weijun Li, Lin Hong, Jörg Conradt, *Senior Member, IEEE*, Jieneng Chen, Zhenyan Zhang, Yiwen Lu, and Alois Knoll, *Senior Member, IEEE*

***Abstract*—Neuromorphic vision sensors such as the Dynamic and Active-pixel Vision Sensor (DAVIS) using silicon retina are inspired by biological vision, they generate streams of asynchronous events to indicate local log-intensity brightness changes. Their properties of high temporal resolution, low-bandwidth, lightweight computation, and low-latency make them a good fit for many applications of motion perception in the intelligent vehicle. However, as a younger and smaller research field compared to classical computer vision, neuromorphic vision is rarely connected with the intelligent vehicle. For this purpose, we present three novel datasets recorded with DAVIS sensors and depth sensor for the distracted driving research and focus on driver drowsiness detection, driver gaze-zone recognition, and driver hand-gesture recognition. To facilitate the comparison with classical computer vision, we record the RGB, depth and infrared data with a depth sensor simultaneously. The total volume of this dataset has 27360 samples. To unlock the potential of neuromorphic vision on the intelligent vehicle, we utilize three popular event-encoding methods to convert asynchronous event slices to event-frames and adapt state-of-the-art convolutional architectures to extensively evaluate their performances on this dataset. Together with qualitative and quantitative results, this work provides a new database and baseline evaluations named NeuroIV in cross-cutting areas of neuromorphic vision and intelligent vehicle.**

***Index Terms*—Neuromorphic vision, distracted driving, advanced driver assistance system, database and baseline evaluations, event encoding, deep learning.**

I. Introduction

IN THE past decade, modern computer vision research has been devoted to conventional cameras .[1] Such kind of cameras have been widely used in robotics, advanced driver assistance system and intelligent vehicles [1], [2]. New algorithms are developed and standardized vision benchmarks are built, which promote rapid development of this field. For example, the growing popularity of deep neural networks [3], [4] in intelligent vehicles and large-scale benchmark such as KITTI [5], Cityscale [6] and ImageNet [7], are interconnected and mutually reinforcing.

However, despite the improved capabilities and breadth of available vision-based systems such as higher resolution and broader dynamic range of cameras, those vision sensors used for intelligent vehicles have remained relatively uniform across platforms in decades. From the first vision-guided Mercedes-Benz robotic van, developed at the Bundeswehr University Munich in 1980s [8], to the most advanced autonomous vehicles developed during DARPA Grand Challenge in 2000s and Google self-driving cars in 2010s [9], conventional cameras are the most important vision sensors. As a result, the algorithms and techniques being designed do not take full advantage of the diverse information provided by modern vision sensors. Since all tasks including perception, localisation, decision-making, and learning are built on top of intelligent sensing, exploring alternative vision sensing approaches instead of tendentious computer vision algorithm developments is of great value that can render subsequent tasks more robust, accurate and complementary.

In this article, we intend to establish a bridge between neuromorphic vision with an innovative Dynamic and Active-pixel Vision Sensor (DAVIS) [10] and intelligent vehicles [11]–[13]. The DAVIS sensor is a bio-inspired neuromorphic vision sensor that works completely differently with conventional cameras. While conventional cameras record the entire intensity images at a fixed rate, DAVIS captures the intensity changes (called *events*) of pixels caused by the motions in a scene asynchronously at the time they occur. This results in a sparse stream of *events* encoding the timestamps, pixel locations, and polarities (sign of brightness changes), as shown in Fig. 1(a) and Fig. 1(b). *Events* are donated as 4-tuple: $e = (x, y, t, p)$, where (x, y) represents the pixel coordinates of events in pixel coordinate system, t records the timestamp when events occur, and p is the polarity of events, which can be either ON $(p = 0)$ or OFF $(p = 1)$ Compared to conventional cameras, neuromorphic vision sensors have several advantages including low energy consumption, high dynamic range (140dB versus 60dB), never suffering from motion blur, and low response latency (microsecond versus millisecond). A neuromorphic vision sensor thus shows a viable alternative or complementary role in the conditions that are challenging for conventional cameras. A corner case in intelligent vehicle is that when the driver is wearing sunglasses, convectional cameras can not capture the state of the driver's eyes, which may result in the failure of the drowsiness driving detection, as shown in Fig. 1(c), Fig. 1(d), and Fig. 1(e). In contrast, DAVIS sensors can sense the blinking of the eyes by penetrating the lens very precisely, as shown in Fig. 1(f) and Fig. 1(g).

Encouraged by the distinctive properties of DAVIS sensors, the rapid development of motion perception research in neuromorphic vision [14]–[18], and the prominent role of distracted driving research in intelligent vehicle, this article will provide a new database and

Manuscript received June 22, 2019; revised February 29, 2020 and July 7, 2020; accepted August 19, 2020. Date of publication October 20, 2020; date of current version February 2, 2022. 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 Advanced Design and Manufacturing for Vehicle Body Open Project under Grant 31815005, in part by the European Union's Horizon 2020 Framework Programme for Research and Innovation (Human Brain Project SGA3) under the Specific Grant 945539, and in part by the Shanghai AI Innovation Development Program 2018. The Associate Editor for this article was C. G. Panayiotou. *(Corresponding author: Guang Chen.)*

Guang Chen is with the School of Automotive Studies, Tongji University, Shanghai 200092, China, also with the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Changsha 410012, China, and also with the Department of Informatics, Technical University of Munich, 80333 Munich, Germany (e-mail: tj_autodrive@hotmail.com).

Fa Wang, Weijun Li, Jieneng Chen, Zhenyan Zhang, and Yiwen Lu are with the School of Automotive Studies, Tongji University, Shanghai 200092, China.

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

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

Alois Knoll is with the Department of Informatics, Technical University of Munich, 80333 Munich, Germany.

Digital Object Identifier 10.1109/TITS.2020.3022921

[1] we name standard frame-based cameras as conventional cameras in this work in order to distinguish from the frame-free neuromorphic vision sensor

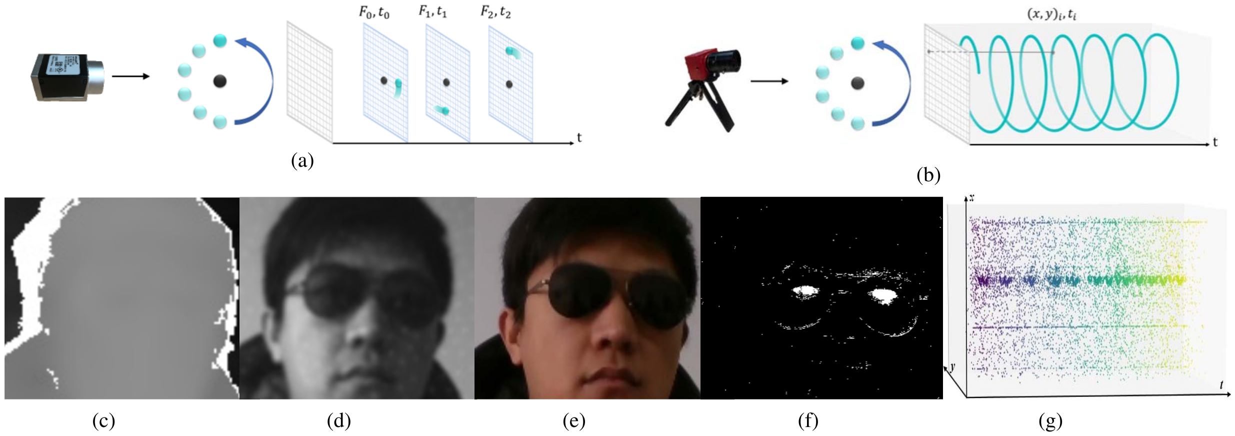


Fig. 1. The differences(top) in the working principle of conventional cameras and neuromorphic vision sensors and a corner case(bottom) in the intelligent vehicle that a driver wears sunglasses while driving demonstrating the super characteristic of DAVIS sensors. A green ball is rotating around a centering black ball. (a) A conventional camera captures all pixel intensities at a fixed frame rate, e.g., black ball and green ball with motion blur. (b) A neuromorphic vision sensor captures only intensity changes caused by the motions of the green ball asynchronously. The images of depth data and infrared data recorded by Realsense D435i are shown in (c) and (d). (e) shows a RGB image recorded by D435i as well that cannot sense what happens behind the sunglasses in these three fashions. (f) and (g) show that a DAVIS sensor can precisely capture eye blinking (and even calculate the blinking frequency) by penetrating the lens.

baseline evaluations that is dedicated to the distracted driving research with an innovative neuromorphic vision sensor. The goal of this work is to serve as a standardized platform for researchers to further investigate the potential usages of neuromorphic vision sensors as a new sensing way in intelligent vehicle.

According to [19]–[22], driver distractions are the leading cause of most vehicle crashes in the word, which result in 40% of driver-related accidents. Distracted driving means that the driver's attention cannot be focused on the main driving task due to manipulation of the buttons in the car, eyes off the road and drowsy driving. Thus, it is very important to sense the driver's head pose, gaze zone, face expression, and hand gestures that infer the level of the driver' distraction. Also, because the proliferation of infotainment systems and secondary driving functionality within a vehicle has led to an environment becoming progressively more distracting for the driver, a well designed human-vehicle interface, that is easy and natural to use, can significantly decrease the distraction of the driver from safety critical operation of the vehicle. Reference [23] reveals that drowsiness driving, driver gaze-zone switching, and driver-vehicle interaction are three major causes of traffic accidents worldwide. Therefore, in this work we focus on the driver drowsiness detection, driver gaze-zone estimation and driver gesture recognition that are three main aspects of distracted driving. We build three datasets recorded with the DAVIS sensor and depth sensor for the distracted driving research that target on the three aspects respectively. To unlock the potential of neuromorphic vision in intelligent vehicles, we adapt state-of-the-art convolutional architectures to the output of the DAVIS sensor, and extensively evaluate the performance of our approaches on the three datasets while comparing with complementary data such as RGB video data, depth data and infrared data. Together with qualitative and quantitative results, our work provides a new database and baseline evaluations named NeuroIV in cross-cutting areas of neuromorphic vision and intelligent vehicles. To the best of our knowledge, this is the first work to connect the neuromorphic vision with the research of distracted driving in intelligent vehicles.

In summary, our contributions are:

- A new database consisting of three sub-datasets with event streams, RGB images, depth data, and infrared data is presented, which is dedicated to driver distraction research on intelligent vehicles: driver drowsiness detection, driver gaze zone detection and driver-vehicle interaction.
- Experiments are designed to explore the potential of the neuromorphic vision sensor in the application of driver distraction recognition. The performances of four baselines and three different encoding methods are provided and compared with the complementary modalities including RGB images, depth data and infrared data.
- All the datasets, the source code, and experimental results of this article will be available online to encourage the comparison of any state-of-the-art methods with our baselines.[2]

We believe that the most valuable characteristic of our work is that it serves as a bridge between neuromorphic vision, modern computer vision techniques and intelligent vehicles, thus bringing the main stream of computer vision based intelligent vehicle research to the attention of neuromorphic vision. We hope that this will gather research interest in new sensing techniques in intelligent vehicles by leveraging the unique strengths of neuromorphic vision sensors, and be a primary starting point that indicates a paradigm shift for bio-inspired visual sensing and perception in intelligent vehicle.

The rest of the paper is organized as follows. In Section II we discuss the related works. In Section III we introduce our NeuroIV dataset in details. In Section IV, we describe four deep learning models utilized in this work. In Section V and Section VI we show the experimental results and explain our findings. In Section VII we conclude our paper and discuss the future directions.

II. Related Works

As a younger and smaller research field compared to conventional computer vision, neuromorphic vision sensors are increasingly receiving attention from computer vision community because they offer advantages over other sensing modalities, such as high temporal resolution, low-bandwidth requirment, low-computational resource requirement and low-latency. In order to exploit potentials of neuromorphic vision sensors in those challenging scenarios for traditional frame-based cameras, many research efforts have been done based on the unconventional output of these sensors. Nowadays, inspired by the inherent advantages provided by neuromorphic vision sensors, they are widely used in real-time interaction systems such as hand gesture recognition [14], [15]. Besides, as natural motion

[2] https://github.com/ispc-lab/NeuroIV

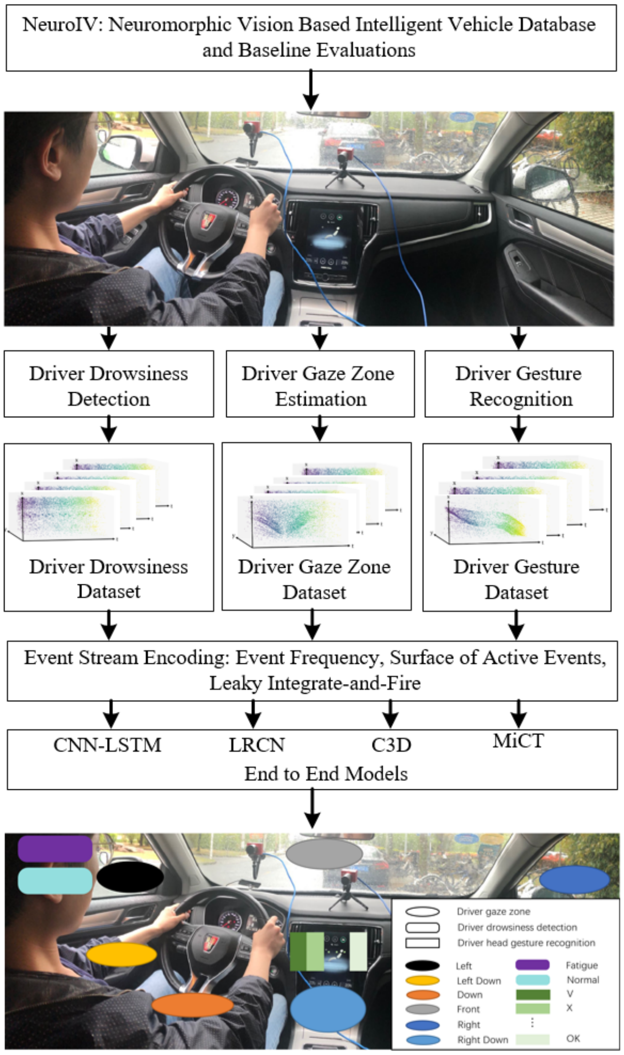

NeuroIV: Neuromorphic Vision Based Intelligent Vehicle Database and Baseline Evaluations
Driver Drowsiness Detection
Driver Gaze Zone Estimation
Driver Gesture Recognition
Driver Drowsiness Dataset
Driver Gaze Zone Dataset
Driver Gesture Dataset
Event Stream Encoding: Event Frequency, Surface of Active Events, Leaky Integrate-and-Fire
CNN-LSTM
LRCN
C3D
MiCT
End to End Models
Driver gaze zone
Driver drowsiness detection
Driver head gesture recognition
Left
Left Down
Down
Front
Right
Right Down
Fatigue
Normal
V
X
⋮
OK

Fig. 2. Framework of the NeuroIV. This work aims at building a bridge of neuromorphic vision with intelligent vehicle. The neuromorphic vision sensor named DAVIS and a depth sensor named RealSense D435i are utilized as the sensing equipment. We focus on the distracted driving in intelligent vehicle with innovative DAVIS sensors where three applications are chosen: driver drowsiness detection, driver gaze-zone estimation and driver hand-gesture recognition. Correspondingly, three datasets are built that consist of the NeuroIV dataset. We then utilize three event encoding methods and four deep convolutional networks that serve as a standardized and open source platform named NeuroIV.

detectors, neuromorphic vision sensors are ideal vision sensor for object recognition [24], object tracking [25], [26], and surveillance and monitoring [27]. With more and more interests and efforts involving in the research community of neuromorphic vision sensors, neuromorphic vision becomes a growing field of research in depth estimation [28], 3D panoramic imaging [29], structured light 3D scanning [30], visual odometry [31] and simultaneous localization and mapping (SLAM) [32].

To be more specific, in vehicle related research, neuromorphic vision sensors have been used to detect, track and estimate the speed of moving vehicles from a stationary view point [27], [33]. It is an ideal choice for intelligent transportation systems to detect and monitor the dynamic objects on road with a neuromorphic vision sensor, since the static background is naturally guaranteed and each cluster of *events* would be a distinct target to locate and track. The first dataset recorded with a on-vehicle neuromorphic vision sensor is the Davis Driving Dataset 2017 (DDD17) [34] that contains approximately 12 hours of annotated driving recordings collected by a car under different and challenging weather, road and illumination conditions. With DDD17 dataset, [18] presents a deep neural network approach on a challenging motion-estimation task: prediction of a vehicle's steering angle. The second large-scale dataset with a on-vehicle neuromorphic vision sensor is from [17] that collects stereo dynamic vision data along with a LiDAR from several types of vehicle platforms, dedicating for future 3D perception tasks [35]. Recently, [36] proposes a network to learn to reconstruct intensity images from event streams directly from data. Their network is able to synthesize high framerate videos (5,000 frames per second) of high-speed phenomena (e.g. a bullet hitting an object) and is able to provide high dynamic range reconstructions in challenging lighting conditions. The above works prove the potentials of neuromorphic visions sensing to be further exploited in vehicle related research fields and application scenarios. Among these, we focus more on driver distraction, since it is a non negligibly hazardous factor in achieving the safe driving goal, and it draws expending attention and efforts from the computer vision community. A lot of techniques have been developed to prevent drivers from possible distractions. In this work, we demonstrate the feasibility of DAVIS applications in such scenarios by assembling three different aspects in distracted driving inside a vehicle, to build up a driver distraction recognition system dedicated for safe driving, which involve driver drowsiness detection, driver gaze zone detection and driver hand gesture recognition.

Driver drowsiness is one of the contributing factors threatening road safety. The process of driver drowsiness generation is a complex issue that is affected by physiological state of drivers and driving environments [37]. Among the many drowsiness driving detection methods, driver facial expressions are regarded as the effective indicators for driver drowsiness. In the early stages of research, many handcrafted parameters such as PERCLOS [38], eye index (EI), pupil activity (PA), and head pose (HP) [39], combined with classifiers are applied to evaluate drivers drowsiness level. Currently, based on the collected frame-based drowsiness driving detection datasets, multifarious convolutional architectures are proposed to improve driving safety. YawDD dataset [40] contains two subdatasets of drivers with various facial characteristics. Each subject in the dataset needs to record three/four videos about mouth movements in different state. Based on the dataset, an efficient and non-invasive approach for detecting driver's yawn was proposed by using a single camera based on long short term memory (LSTM) networks. FI-DDD dataset [41] contains 18 drivers, there are 450 video clips recorded in day and night for training and testing. A long-term multi-granularity deep framework was proposed to detect driver drowsiness. A public NTHU-DDD dataset provided by [42] is consisted of 36 subjects of different ethnicities. The driver's behaviors are recorded with and without glasses/sunglasses in various simulated driving scenarios. A novel hierarchical temporal deep belief network (HTDBN) method was proposed for drowsy detection.

Driver gaze zone provides an approximate estimation of drivers' eye-gaze direction and thus can be regarded as an attention indicator. From it, we can better understand the driver's intended moves and thus it enables a system for safer driving experience. A straightforward way to estimate gaze zone detection is to discretize the front view space into several sections and turn it into a classification task. Extracted features from driver's visual characteristics can be then fed into algorithms such as support vector machine (SVM), random forest, deep learning networks or etc, outputting the current or impending area of attention focus. Previous works mainly uses an RGB camera to monitor the drivers from the front, with different techniques in data processing. Reference [43] collected 300,000 frames

TABLE I
NEUROIV DATASET DESCRIPTION

Datasets		Classes	
Driver Drowsiness	1. Normal driving	2. Drowsiness driving	
Driver Gaze-Zone	1. To the left 4. To center stack	2. To left stack 5. To the right	3. Switch downward
Driver Hand-Gesture	1. Wiper 4. MoveLeft 7. PushDown 10. OK 13. V 16. OneTap	2. MoveUp 5. MoveRight 8. PushLeft 11. No 14. Pinch	3. MoveDown 6. PushUp 9. PushRight 12. X 15. Expand

from 12 subjects, extracts edge information from images and proposes the method of matching with face template and an SVM classifier to roughly estimate the drivers' gaze zone. Reference [44] predicts gaze zone from head pose estimation based on pixel intensities through joint classification and regression with a multi-loss convolution neural network trained on 300W-LP [45]. Reference [46] utilizes a Siamese architecture and they proposes a novel loss function to improve the learning result. The regression network layers are trained on BIWI [47] and 3dHP [48]. Some others put forward solutions of Long Short-Terms Memories modules, using a Recurrent Neural Network [49] to aggregate information over time, and therefore the prediction and estimation is made upon a series of frames. Also, some other works tackle the problem by utilizing traditional expert systems with if-then-else logic [50]. And more recently, end-to-end deep learning techniques with convolution networks are proved to be very effective, like in [51] which provide a dataset of 11 drives from 10 subjects for model training.

The human-computer interaction system improves the driving experience, but it also distracts the driver's attention. There are many in-vehicle systems that can recognize the driver's hand gestures, thereby eliminating the need for the driver's physical operations to accomplish some tasks, such as adjusting the volume of the music and the temperature of the air conditioner. There are many previous works on hand gesture recognition. In [52], a RGB dataset consists of 9 command gestures for Human-Robot Interaction (HRI) was proposed. In [53], RGB-D video sequences comprised of more than 100K frames of 45 daily hand action categories were collected, involving 26 different objects in several hand configurations. To obtain hand pose annotations, mo-cap system are used that automatically infers the 3D location of each of the 21 joints of a hand model via 6 magnetic sensors and inverse kinematics. In [14], the first event-based hand gesture dataset (DvsGesture) was collected by using DVS, it comprised 11 hand gesture categories from 29 subjects under 3 illumination conditions. To achieve hand gesture recognition, there are many detection algorithms have been developed by using improved Neural Networks. Reference [54] proposed a method using Convolutional Neural Networks (ConvNets) through bidirectional rank pooling and adopts Convolutional LSTM Networks (ConvLSTM) to learn long-term spatiotemporal features from short-term spatiotemporal features extracted using a 3D convolutional neural network (3DCNN) at body and hand level. Reference [52] analyzed a Convolutional Long Short-Term Memory Recurrent Neural Network (CNNLSTM) in the context of gesture recognition, in order to show that CNNLSTM outperforms both plain CNN and LSTM in gesture recognition. Reference [55] proposed a postprocessing framework based on spiking neural networks that can process the events received from the DVS in real time, and provided an architecture for future implementation in neuromorphic hardware devices.

III. THE NEUROIV DATASET

A. Data Collection

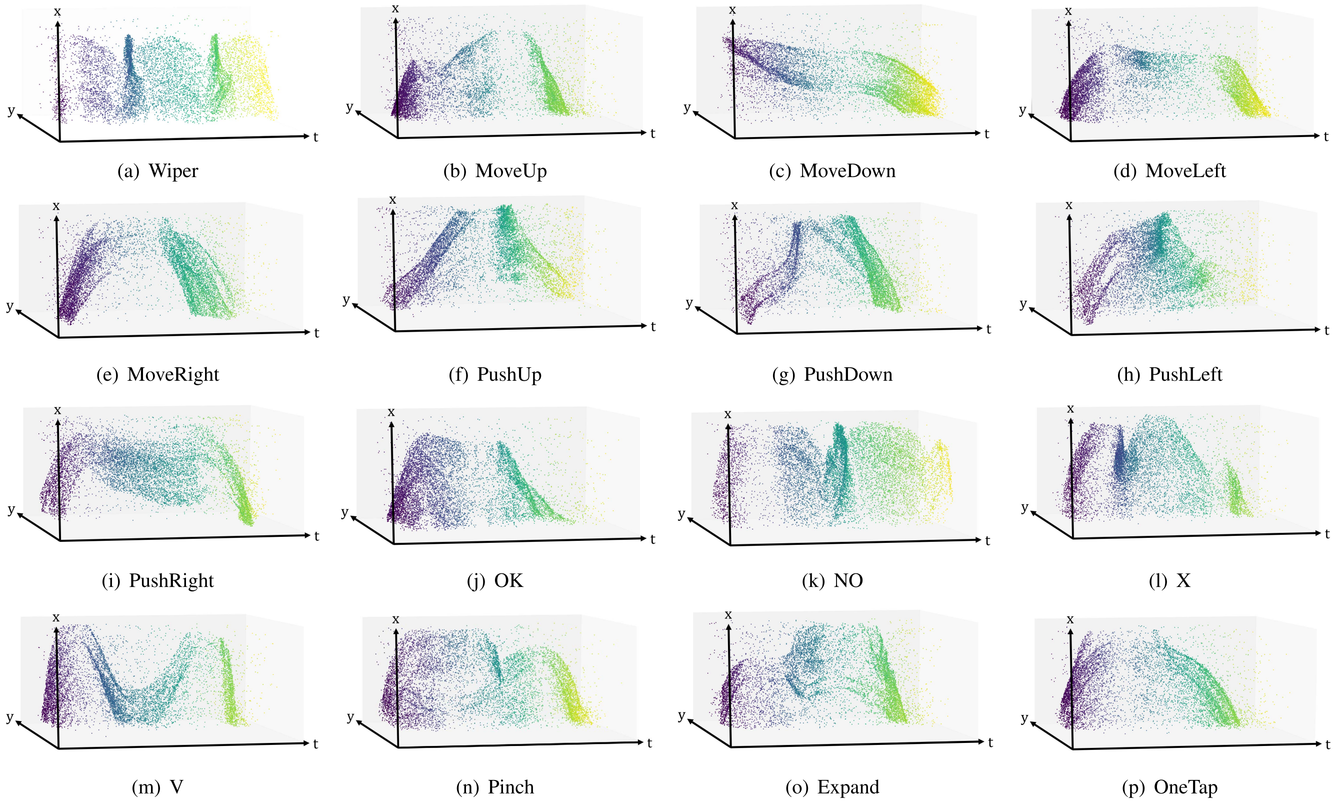
To build the NeuroIV dataset, the DAVIS346[3] is used as our main recording sensor, which has a resolution of 346 × 260 pixels, a temporal resolution of 1 us, and an outstanding dynamic range (up to 140 dB). Additionally, as a complement of our dataset, conventional RGB data, infrared data and depth data are recorded simultaneously by a RealSense D435i with a resolution of 640 × 360 pixels. Two sensors are placed on top of the dashboard to simulate the monitoring process of the driver's face motions and the driver's hand gestures. There are 30 subjects (of whom, 3 are women and 27 men) comprised of campus guards, graduates and college students, participated in the dataset collection. All the subjects have the Chinese C1 driving license. Their ages range from 20 to 50, and the average age is 26. All of them are in good health and have enough sleep before data collection. A video tutorial for Driver Gaze-Zone dataset and Driver Hand-Gesture dataset or a text one for Driver Drowsiness dataset, seen in Table III, is playing in loop mode that teaches the subjects how to perform each of pre-defined behaviors precisely. Before starting the recording, we also give a short description of all pre-defined behaviors including the names and the meanings of them. Thus, the subject is guided the whole recording process accordingly to ensure the high quality of each recording. However, the distance between subjects and the sensors are varied in a limited range (up to 15cm) to mimic the realistic driving environment. To ensure the safety of the subjects, all the data are recorded in a simulated environment.

B. Dataset Statistics

The NeuroIV dataset consists of three sub-datasets: Driver Drowsiness dataset, Driver Gaze-Zone dataset and Driver Hand-Gesture dataset. A summary of NeuroIV dataset statistics is shown in Table I and Table II. A visualization of samples from the Driver Hand-Gesture dataset is shown in Fig. 3. We describe them separately as follows.

1) Driver Drowsiness Dataset: There are two different driving states including normal driving and drowsiness driving designed in Driver Drowsiness dataset. We use the DAVIS346 and the RealSense D435i placing on the dashboard to record drivers' drowsiness state. In normal driving state, the driver blinks at a regular frequency and doesn't yawn or doze off. In drowsiness driving state, the driver blinks frequently, and appears to yawn and doze off in the process of driving. The dataset is recorded in daytime with nine scenarios shown in Table III, each of which of every subject is recorded twice, to imitate the real driving environments more comprehensively. There

[3] https://inivation.com/dvs/dvs-product-variants/



(a) Wiper (b) MoveUp (c) MoveDown (d) MoveLeft

(e) MoveRight (f) PushUp (g) PushDown (h) PushLeft

(i) PushRight (j) OK (k) NO (l) X

(m) V (n) Pinch (o) Expand (p) OneTap

Fig. 3. Visualization of *event* stream from random samples of Driver Hand-Gesture Dataset. Each sub-figure corresponds to one sample.

TABLE II
NEUROIV DATASET STATISTICS

Datasets	The numbers of Sample					Descriptions		
	DVS	RGB	Depth	Infrared	Subjects	Scenes	Task	Labels
Driver Drowsiness	540	540	540	540	30	barefaace,glasses,sunglasses	classification	2
Driver Gaze-Zone	1500	1500	1500	1500	30	-	classification	5
Driver Hand-Gesture	4800	4800	4800	4800	30	-	classification	16

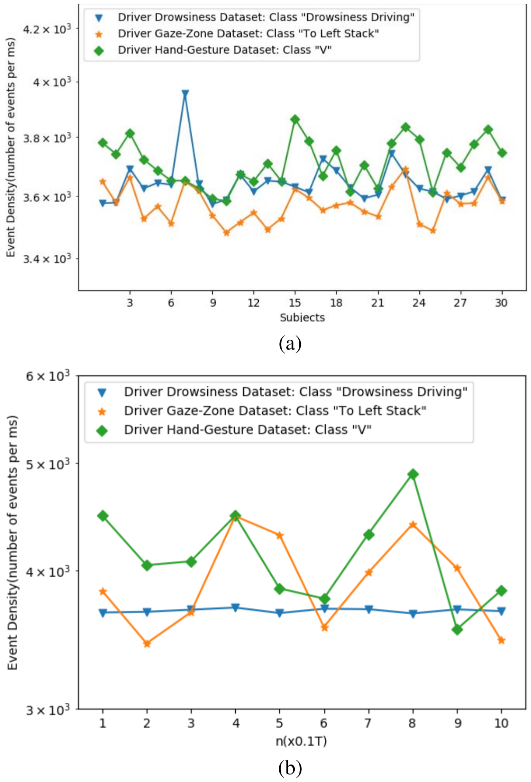
TABLE III
TEXT TUTORIALS FOR DRIVER DROWSINESS

Normal Driving	Drowsiness Driving
1. Driver with bareface,looking ahead	5. Driver with bareface, looking ahead with yawn
2. Driver with bareface, looking at the rearview mirror occasionally	6. Driver with bareface, looking at the rearview mirror with yawn
3. Driver with glasses, looking ahead	7. Driver with glasses, looking ahead with yawn
4. Driver with sunglasses, looking ahead	8. Driver with sunglasses, looking ahead with yawn
	9. Driver with bareface, looking ahead with yawn even nod

are a total of 2160 samples (540 for each data modality) in this dataset. Details are described in Table I and Table II.

2) Driver Gaze-Zone Dataset: The driver gaze-zone switching actions serve as the indicators of driver's attention shifting behaviors. While we take account of the left and right zone switching as the signal of attention shifting to the driver's sides, the looking downward action (at the instrument panels) or nodding is also captured as gaze switching to the lower front zone. Besides, considering occasional gaze attention shift to the storage stack in left side and central panel in right side, we also include the switching action to the lower left zone and to the lower right zone. In short, there are five driver gaze-zone switching actions: to the left, to the right, switching downward, to the central stack and to left storage stack. Subjects are asked to sequentially perform these actions for 10 times. In total, there are 6000 samples, 1500 for each modality collected from 30 subjects. Details are described in Table I and Table II.

3) Driver Hand-Gesture Dataset: Gesture-based contactless interface can significantly decrease the distraction of the driver from the interaction with infotainment systems and secondary driving functionality. In this dataset, there are 16 kinds of gestures designed for the specific secondary driving functionalities such as *Pinch* for *ScreenZoomIn* and *Expand* for *ScreenZoomOut*. Each subject performs 10 times of all the gestures under natural sunlight conditions respectively. In total, we have 19200 samples, 4800 for each data



(a)

(b)

Fig. 4. (a) Subjects versus average *event* densitiy. Each line corresponds to one specific class in NeuroIV dataset. (b) The variation of average *event* density over time. Each line corresponds one specific class in NeuroIV dataset.

modality. Fig. 3 shows the visualization of *event* streams from samples of 16 classes gestures. Details of the dataset are described in Table I and Table II.

C. Dataset Characteristics

1) Inter-Subject Diversity: Due to individual differences among the subjects, the same class of driver's behaviors can be performed in different ways. In order to represent variation among different subjects with the same behavior, event density dots of the same scenario under natural sunlight condition with same color and shape are ploted in Fig. 4(a). As shown in Fig. 4(a), the x-axis and y-axis represent the identifier numbers of 30 subjects and the average *event* densities (number of *events* per ms) of all samples from a selected class, respectively. For each sub-dataset, we select one class of driver's behaviors and calculate the average *event* densities of all the samples from every subject, which correspond to 30 points as shown in Fig. 4(a). It is clear to see that the *event* densities across different subjects varies in the range of 3000 events/ms to 4000 events/ms. For the class "To the Left" in Driver Gaze-Zone dataset, "V" in Driver Hand-Gesture dataset and "Drowsiness Driving" in Driver Drowsiness dataset, each pair of their *event* densities has little difference. However, because of various ranges of motion across subjects, the average *events* density varies, which makes the NeuroIV dataset more challenging with higher diversity.

2) Variation Over Time: Along the time dimension, *event* density varies significantly because of the changing speed and magnitude of the motion patterns of different driver's behavior in NeuroIV dataset. In Fig. 4(b), we randomly choose a subject and show the variations of *event* densities over time. The x-axis of Fig.4(b) represents ten sampling points along the period of each sample, and the y-axis represents the average *event* density of different samples from the selected subject. Each curve in Fig. 4(b) corresponds to one representative class in NeuroIV dataset. It can be seen that the class "V" in Driver Hand-Gesture dataset and the class "To the Left" in Driver Gaze-Zone dataset seem to be more significantly varying behavior than the rest one. It is not surprising that the curve of the class "Drowsines Driving" in Driver Drowsiness dataset is relative flat because the relatively vigorous motions, such as blinking and yawn, happen randomly and repeatable in one period.

3) Inter-Class Diversity: Fig. 5 shows the *event* density variation among different classes of the NeuroIV dataset, where each colored dot corresponds to one subject of all his/her recorded samples of a certain class. The x-axis of Fig. 5 represents classes of the three sub-datasets in NeuroIV dataset, and the y-axis represents the average *event* density. The cyan triangles indicate the mean of average *event* densities of 30 subjects for each class. As can be observed, a relatively small variation is existing in the Driver Gaze-Zone dataset and Driver Hand-Gesture dataset as the classes in these two datasets are following a sequence of pre-define motion patterns while the motion patterns of drowsiness driving such as eye blinking and yawning occur randomly.

IV. Event Data Encoding and Deep Learning Models

Since the output of a DAVIS sensor is an asynchronous stream of *events* that is fundamentally different with images of conventional cameras, existing computer vision algorithms cannot be directly applied to it. We first investigate different *events* data encoding methods by mapping an *events* slice with time interval T into an image-like 2D representation prior to any learning models. Next, we briefly describe four state-of-the-art convolutional architectures adopted to evaluate our encoding methods on the NeuroIV dataset.

A. Event Data Encoding

In this section, three different encoding methods [56]–[58], named *Event Frequency (FRQ)*, *Surface of Active Events (SAE)* and *Leaky Integrate-and-Fire (LIF)* are utilized for converting *event* streams into event-frames.

1) Event Frequency: Knowing that the occurrence frequency of a certain *events* within a given time interval could be an evidence of whether they are valid *events* or just noise, we could tally the occurrence of the *events* at each pixel (x, y). The proposed range normalization equation inspired by [59] is shown below:

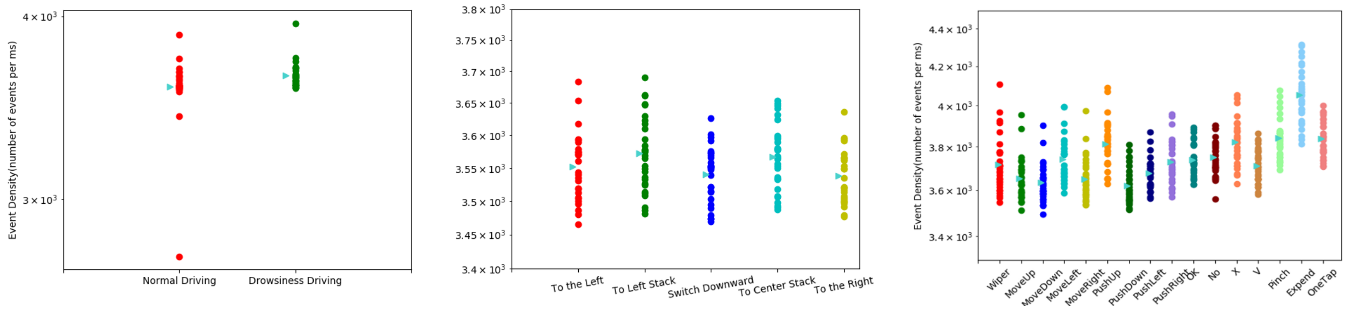
$$\sigma(x, y) = 255 \cdot (2 \cdot \frac{1}{1 + e^{-n(x,y)}} - 1) \tag{1}$$

where $n(x, y)$ is the total number of the occurred *events*, $\sigma(x, y)$ is the pixel value of the event-frame. To fit the 8-bit image, we normalize the range of $\sigma(x, y)$ to be between 0 and 255, orderly. For the *Event Frequency* encoding method, the edges of moving objects will be strengthened to a great extent, which is beneficial for detection and recognition tasks as event-frames have a clear profile of objects.

2) Surface of Active Events: The *Surface of Active Events (SAE)* [56] approach is applied to reflect time information while the pixel value and its gradient can tell the moving direction and speed of the *event* stream. Specifically, each incoming *events* $[t, x, y, p]$ will change the pixel value t_p at (x, y) according to the time-stamp t. In this way, an grayscale image frame is acquired according to the time-stamp of the most recent *events* at each pixel:

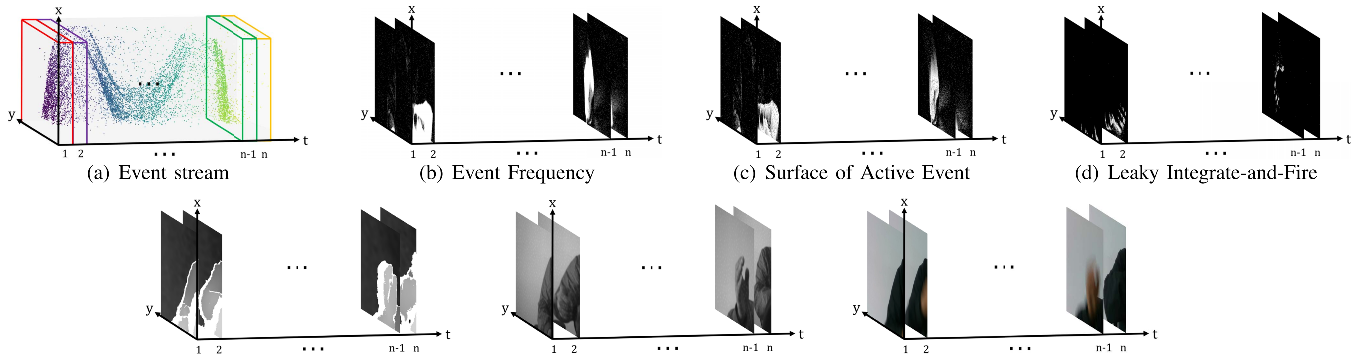
$$SAE : t \Rightarrow t_p(x, y) \tag{2}$$

We conduct numerical mapping to get an 8-bit grayscale image where the pixel value $\sigma(x, y)$ is calculated by $t_p(x, y)$ and $t_0(x, y)$



(a) Driver Drowsiness Dataset (b) Driver Gaze-Zone Dataset (c) Driver Hand-Gesture Dataset

Fig. 5. Event density variation among different classes of the NeuroIV dataset. Each colored dot corresponds to one subject of all his/her recorded samples of a certain class. The x-axis of represents classes of the three sub-datasets in NeuroIV dataset, and the y-axis represents the average *events* density. The cyan triangles indicate the mean of average *events* densities of 30 subjects for each class.



(a) Event stream (b) Event Frequency (c) Surface of Active Event (d) Leaky Integrate-and-Fire

(e) Depth frames (f) Infrared frames (g) RGB frames

Fig. 6. (a). An event stream is split into several event slices. Each event slice corresponds to a cube cell in the image. (b) Event-frames converted from event slices by *Event Frequency*. (c) Event-frames converted from event slices by *Surface of Active Event*. (d) Event-frames converted from event slices by *Leaky Integrate-and-Fire*. (e)Frames generated from depth data modality video. (f)Frames generated from infrared data modality video. (g)Frames generated from rgb data modality video.

(initial time-stamp) as follows:

$$\sigma(x, y) = 255 \cdot \frac{t_p(x, y) - t_0(x, y)}{T} \quad (3)$$

3) Leaky Integrate-and-Fire: According to the LIF (Leaky Integrate-and-Fire) neuron model [57], each neuron has its very own Membrane Potential (MP). Such an MP could be influenced by time-lapse or input spikes. If the MP exceeds the preset threshold, a firing spike output would be generated. The encoding procedure of the LIF neuron model is shown in Fig. 7.

We regard each pixel (x, y) in a time interval T as a neuron with its MP and firing counter as n. Each *events* at (x, y) would cause a step increase of the pixel's MP; and simultaneously each pixel's MP will decay at a fixed rate. We count the number of the firing spike outputs of pixel (x, y) when MP exceeds the threshold into a firing counter $n(x, y)$. After each time interval T, the firing spike counter n will be reset to 0 and starts the counting in the next time interval again. Following the same normalization mechanism as the encoding method *Event Frequency*, we get the final 8-bit grayscale event-frame.

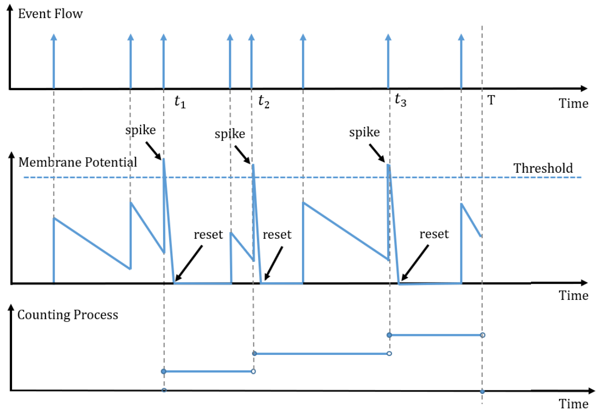


Fig. 7. Coding mechanism of the LIF neuron model. **Top** is an asynchronous *event* stream flow. **Middle** is the Membrane Potential (MP). At time t_1, t_2 and t_3, there is a LIF neuron spike, and simultaneously MP will decay at a fixed rapid rate to reset. **Bottom** is the counting process, each spike is counted when MP exceeds the threshold. After each time interval T, the firing spikes counter n will be reset to 0 and starts the counting in the next time interval.

B. Event Data Visualization

To better understand how DVS sensors manage the movements and illustrate why our experiment settings make sense, we visualize raw data of DVS event slices and the encoded event-frames in Fig 8 and Fig 9. Sub-figures are presented in time-windows of 10ms, 30ms and 50ms from left to the right. As we extend the time-window, the event slices would first become more recognizable as edges and shape enriched by more accumulated events. But wider time-windows also come with more noise, which could result in overexposure. The three different encoding methods also have different effects on event-frames. As we can see from Fig 8 and Fig 9, the *Event Frequency* approach provides the clearest images of the event frames. The *Surface of Active Event* approach reflects the motion of the

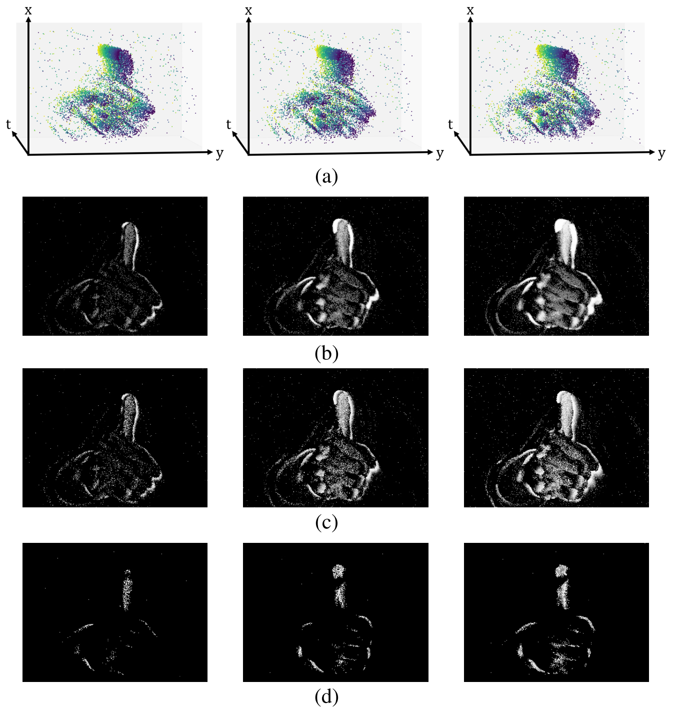


Fig. 8. (a) Visualization of *event* slices (samples from the class “OK” in Driver Hand-Gesture dataset) with different time-windows: 10ms, 20ms, 50ms, from left to right. (b) Encoded event-frames of *event* slices by *Event Frequency*. (c) Encoded event-frames of *event* slices by *Surface Active Events*. (d) Encoded event-frames of *event* slices by *Leaky Integrate-and-Fire*.

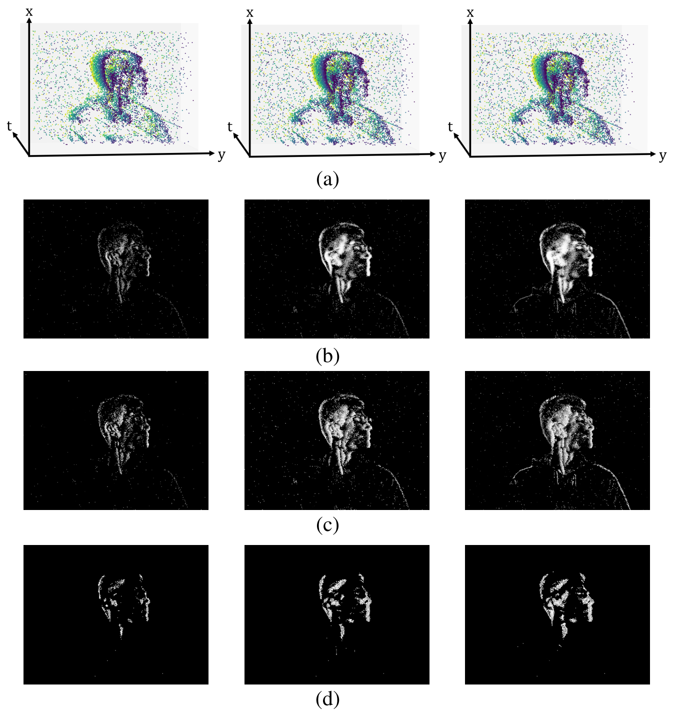


Fig. 9. (a) Visualization of *event* slices (samples from the class “To the Left” in Driver Gaze-Zone dataset) with different time-windows: 10ms, 20ms, 50ms, from left to right. (b) Encoded event-frames of *event* slices by *Event Frequency*. (c) Encoded event-frames of *event* slices by *Surface Active Events*. (d) Encoded event-frames of *event* slices by *Leaky Integrate-and-Fire*.

moving objects. The *Leaky Integrate-and-Fire* approach emphasize the active events and inhibit the noise events such as events from the background.

C. Deep Learning Models

We adapt four state-of-the-art convolutional architectures to the encoded output of the DAVIS sensor. As these deep learning models show great performance on video-level perception tasks such as action recognition [60] and gesture recognition [61] with conventional cameras, we are interested in how these convolutional networks perform on NeuroIV dataset. The network architectures and implementations are briefly described as follows.

1) CNN-LSTM: The CNN-LSTM is a popular architecture widely used in the video data recognition tasks [62]. The model is composed of a pretrained Inception V3 module [63] for feature extraction, and a LSTM module [64] for prediction. The Inception V3 is pretrained on ImageNet, and output the feature vectors with the length of 2048. The features from the Inception V3 will be stored, and then fed to a LSTM network with 2048 dimensional hidden states. The LSTM is followed by two fully connected layers. To prevent over-fitting, dropout layers of 0.5 are placed after each fully connected layer. After the LSTM module, The tensors after the LSTM module are flattened to one-dimensional vectors. The two fully-connected layers are with output the length of 512 and class_num respectively, where class_num stands for the total number of class. Each FC layer is followed by a ReLU non-linear activation function. A softmax layer is added after FC layers to output the probability of predicted class. For feature extraction, we extract N frames in the pretrained Inception V3 module. N is set to 40 for Driver Gaze-Zone dataset and Hand-Gesture dataset while N is set to 200 for Driver Drowsiness dataset. To sum up, we firstly use a pretrained extractor to obtain CNN visual features, and then use these visual features to train a LSTM network to make the final prediction.

2) LRCN: The LRCN is a Long-term Recurrent Convolutional Network (LRCN) architecture proposed in [65]. The LRCN network is composed of visual and sequential components, i.e. a CNN visual module and a LSTM sequential module. A time distributed CNN visual module contains N CNN blocks. A CNN block is designed as a VGG-16 block [66]. Each frame is extracted to a feature map by a separate CNN block. N feature maps are flattened to N feature vectors, and then together fed to a LSTM network (256 cells) followed by a FC layer with length of class_num with the softmax activation to make the final prediction [65]. The setting of N is the same with CNN-LSTM. Each 346 × 260 DVS frame or 640 × 360 image from Realsense D435i raw input is resized into a 80 × 80 frame as the input of the time distributed CNN visual module. Under the proposed system, the parameters of the model's visual and sequential components can be jointly optimized by maximizing the likelihood of the ground truth outputs. It is worth mentioning that the network is trained in an end-to-end fashion compared to the introduced CNN-LSTM network before.

3) C3D: The C3D is a widely used architecture for 3D video recognition [67]. To fully exploit discriminative spatial temporal features in video recognition, we adopt a 3D convolutional network following the architecture proposed in [67]. It contains 8 3D convolution layers, 5 pooling layers, 3 fully connected layers, and a softmax layer. To prevent over-fitting, dropout layers of 0.5 are placed after each fully connected layer. The 3D feature map can be flattened to one dimensional vectors by fully-connected layers. Three fully-connected layers have the output length of 4096, 4096 and N_{output} respectively, where N_{output} stands for the total number of class. Each fully connected layer is followed by a ReLU non-linear activation function, except for the last layer that is followed by a softmax function instead. During the training phase, all layers in the network are trained from scratch. An *event* stream is cut into a sequence of *event* slices that are then converted into event-frames using three different encoding methods mentioned in Section IV-A. The input 3D spatio-temporal volume contains 40 resized frames (200 frames for drowsiness detection dataset). The final Softmax classifier outputs the probabilities of predicted *event* stream class.

4) MiCT: Mixed 2D/3D Convolutional Tube (MiCT) net architecture as described in [68] integrates 2D CNNs with 3D convolution layers. It enables 3D CNNs to extract deeper spatio-temporal features with fewer 3D spatio-temporal fusions and thus reduces the complexity that a 3D convolution needs to encode at each round of spatio-temporal fusion. The MiCT architecture uses five 3D convolutions, one at the entrance of the network and one at the beginning of each of the four main ResNet blocks. After each 3D convolution, features of the two branches are merged with a cross domain element-wise summation. Experiments are based on the 18 layers version of the ResNet backbone. The temporal stride is 16 and the spatial stride is 32. The first 3D convolution has a temporal stride of 1 to fully harvest the input sequence. Weights are initialised from ImageNet pre-trained weights. For 3D-ResNet, the 3D filters are bootstrapped by repeating the weights of the 2D filters N times along the temporal dimension, and rescaling them by dividing by N. Data input details are the same as the previous three.

V. Experiments

In this section, we evaluate the performance of the state-of-the-art convolutional architectures with proposed encoded methods on NeuroIV dataset. We focus on the distracted driving research with DAVIS sensors. The goal of this work is to provide a new database and baseline evaluations instead of developing advanced approaches to get the best performance on our dataset. By analyzing experimental results, we demonstrate the potential of the neuromorphic vision as a new sensing techniques in intelligent vehicles with the help of modern computer vision techniques. In details, we design our experiments to investigate the following three questions: how to design a better encoding method, what are the performance of the convolutional networks on the given tasks and what are the performance of different data modalities on the given tasks, even if the networks are pre-trained on frames collected by convertional camera. Regarding the first question, we analyze the performance of the networks over different event-encoding time-windows in Section V-A and different encoding strategies in Section V-B on a subset of NeuroIV dataset. Regarding the second question, we conduct an extensive study by four trained deep learning models over the entire NeuroIV dataset in Section V-C. Concerning the last question, we conduct experiments by four models as mentioned above over four data modalities on three subsets in Section V-D.

A. Sensitivity Analysis of Different Event-Encoding Time-Windows

In this section, we investigate the effects of the event-encoding time-window on the performance of deep learning models with NeuroIV dataset. Several visualizations of *events* slices in 10 ms, 30 ms, 50 ms, and 70 ms are shown in Fig. 8(a). These encoding time-windows are chosen to be in the range of an equivalent frame-rate of 15 fps to 100 fps. It can be observed that the longer the encoding time-window, the larger is the blur effect appearing at the boundary of the objects. The reason is that objects move a longer distance under the image pixel coordinates during a longer time-window. We hypothesize that there is a trade-off between the length of the encoding time-window and the discrimination power of motion cues in the *event* slice. In other words, a certain time-window is existing that can be effectively exploited by models to provide a better performance than others.

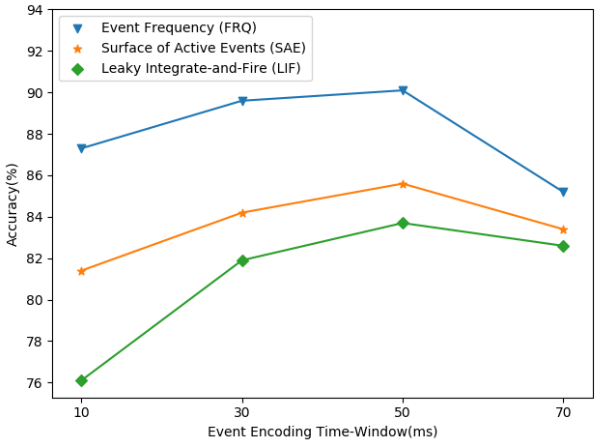


Fig. 10. The recognition accuracy by CNN-LSTM model on a subset of Driver Hand-Gesture dataset with respect to different lengths of event-encoding time-window.

TABLE IV

Accuracy Comparison of Different Encoding Methods

Encoding Method	Accuracy(%)			
	10ms	30ms	50ms	70ms
Event Frequency	87.3	89.6	90.1	85.2
Surface of Active Events	81.4	84.2	85.6	83.4
Leaky Integrate-and-Fire	76.1	81.9	83.7	82.6

It is understandable that fewer *events* will been encoded into event-frames (regardless of the encoding methods) when the encoding time window is small. Since less information is embedded in the *event* slice, the event-frames can only reflect relative small motion that lose the power of discrimination. In contrast, with a long encoding time-window, the blur disrupts the clear boundary of objects, in particular when the object motion such as fingers' motions are relatively high. Consequently, the encoded event-frames fail to capture the sharpness of moving edges that could be necessary for motion related recognition tasks such as hand gesture recognition.

Fig. 10 reports the quantitative results of our CNN-LSTM network with four encoding time-windows and three different encoding strategies. As can be observed, the best performance is achieved with the encoding time-window of 50 ms with all of the three encoding methods. The performance declines in smaller or larger encoding time-window. Therefore, we set the encoding time-window to the best value, 50 ms in the following experiments.

B. Comparison on Different Encoding Methods

After setting the encoding time-window to 50 ms, we are interested in evaluating the performance of different encoding methods on the given tasks. Experiments are done with CNN-LSTM models on a subset of Driving Hand-Gesture dataset that is the same with Section V-A. The results are shown in Table IV.

It is interesting to notice that the encoding method *Event Frequency* is slightly better than the other two. This is to be expected, since *Event Frequency*, strengthening the profile of objects, can therefore maximize the discriminability of the encoded frames, while at the same time, restrain the non-ideal effects from events noises that the encoded frames can be modeled to better extract meaningful information. Regarding the encoding method *SAE*, although the raw timestamp information is directly utilized as the pixel value and its gradient can reflect the moving direction and speed of the *event* stream, the main problem is the inability to filter out noise *events* due to

TABLE V

EXPERIMENTS RESULTS ON THE NEUROIV DATASET OF DVS. THE THIRD COLUMN SHOWS THE RECOGNITION ACCURACIES (ACC) OF CORRESPONDING DATASETS AND CONVOLUTIONAL ARCHITECTURES

Datasets	Convolutional Achitecture	Acc (%)
Driver Drowsiness	CNN-LSTM	63.54
	LRCN	55.61
	C3D	73.47
	MiCT [69]	60.60
Driver Gaze-Zone	CNN-LSTM	77.09
	LRCN	91.65
	C3D	96.98
	MiCT	75.20
Driver Hand-Gesture	CNN-LSTM	63.61
	LRCN	73.74
	C3D	89.02
	MiCT	63.30

the ignorance of *event* occurrence frequency. While the *LIF*-based method set a stricter rule for *event* encoding, it leads to less noise but correspondingly increases the loss of useful information. A visualization of three encoding methods is in Fig. IV-B. We can see that the encoding method *SAE* (Fig. 8(c)) produces a long blur-tail at the boundary of the fingers, the encoding method *LIF* (Fig. 8(d)) capture very sparse *event* information, while the encoding method *Event Frequency* (Fig. 8(b)) makes a balance between noise elimination and valid information encoding. In the following experiments, we choose the encoding method *Event Frequency* for the entire NeuroIV dataset evaluation.

C. Results on the Entire NeuroIV Dataset

In this section, we present quantitative results on the entire NeuroIV dataset with state-of-the-art convolutional networks. There are three sub-datasets in our NeuroIV dataset, which correspond to three independent experiments focusing on driver drowsiness detection, driver gaze-zone recognition and driver hand-gesture recognition, respectively. In each experiment, all sample data are randomly divided into three parts based on the subject IDs: training data, validation data, and testing data, with a ratio of 6:2:2. The subjects are separated with no overlapping in the three parts. The event-encoding method is *Event Frequency*, and the encoding time window is fixed to 50 ms. The implementation details including training parameters settings are described in Section IV-C.

Table V shows our experimental results. We can see that the best accuracy of three tasks (driver drowsiness detection, driver gaze-zone recognition and driver hand-gesture recognition) with DVS sensors are 73.47%, 96.98%, and 89.02%, respectively, which are close to state-of-the-art performance with no fine-tuning of our model parameters. It is worth noting that our goal of these experiments is to show that, by only relying on event-frames without fusing other modality data, state-of-the-art convolutional networks are able to achieve great performance on these given complex tasks. This means that neuromorphic vision sensor is suited to motion-related tasks such as gesture recognition and classification. The experimental results serve as strong baselines for the NeuroIV. Furthermore, the CNN architecture of CNN-LSTM model is trained on frames collected by standard frame-based cameras that show it is possible to leverage transfer learning from pre-trained convolutional networks. In addition, the performance of the driver drowsiness detection is worse than the other two. This is to be expected, since the average length of a data sample is 60s, we only extract features from the sequence of 200 event-frames in each sample as input to the CNN architecture. The performance driver gaze-zone estimation is better than the other two. Given the large-scale of the driver gaze-zone estimation dataset and its distinguishable classes compared with the driver hand-gesture interaction dataset, it is explainable as large amount of data and the small number of classes make the network training to be an easier task to model the statistics of the data.

TABLE VI

COMPARISON OF ROBUSTNESS OF DVS AND RGB

Datasets	Data Format	Convolutional Achitecture	Illumination	Acc (%)
Drive Hand-Gesutre	DVS	C3D	Normal	89.0
			Low	61.0
	RGB	C3D	Normal	94.5
			Low	20.4

D. Comparison With Other Modalities

In this section, we present quantitative results on the entire NeuroIV dataset with four convolutional networks on four different modalities: event frames, RGB images, Depth images, and Infrared images. Table VII shows the comparison results. For the driver drowsiness dataset, the C3D with RGB images get the best performance (accuracy with 81.6%). The event frames with C3D is up to the second (accuracy with 73.5%). For the driver gaze-zone dataset, the C3D with infrared images get the best performance (accuracy with 98.4%) while the C3D with event frames and RGB images are up to second (accuracy with 97.0%). For the driver hand-gesture dataset, the performances of RGB images, depth images and infrared images are close to each other (accuracy with 94.5%, 96.1% and 95.2%, respectively). Base on the experimental results, we could draw two conclusions. Firstly, the input modality of RGB images has the best performance among all of them. It is expected as these convolutional networks are trained with RGB images. Secondly, among the performance with event frames, depth images and the infrared images, experimental results with the event frames get slightly better performance with the driver drowsiness dataset than with the depth images and infrared images. For the driver gaze-zone dataset, the experimental performance of the event frames is only slightly worse than the infrared images. It indicates that the event frames with a simple encoding method *Event Frequency* can achieve even better performance than the depth images and infrared images.

Table VIII shows the time consumption on the NeuroIV dataset when using various types of approaches and data modalities. We indirectly measure the models' complexities by comparing their inference time consumption. Among the three methods, C3D consumes the most inference time for highest accuracy, while CNN-LSTM performs slight better than LRCN. When comparing different data modalities, RGB data and DVS event data have very similar computation time. RGB data is only with subtle superiority.

E. Robustness and Adaption Advantage Over RGB Sensors

To further unlock and present the potential of DVS based modality, we have acquired an extended dataset in a very dimmed lighting condition with all data collection settings the same and compared DVS and RGB modalities using C3D approach and driver hand gesture dataset in normal and low illuminations to verify the robustness and adaption advantage of DAVIS over RGB sensors. As can be seen in Table VI, DVS modality shows better robustness against such challenging ambient lighting. In this mini-demonstration, RGB based approach suffers from a drastic performance drop of 74.1%, while DVS based approach is only affected by 28%. Although lighting condition is always a major challenge factor in passive sensor based approaches, DVS based approach is much robust and adaptive than traditional RGB snesors in such extreme conditions. DVS sensors are also much tolerant of extremely bright lighting and rapid illumination

TABLE VII
COMPARISON OF FOUR DATA TYPES ON THREE SUB-DATASETS

	CNN-LSTM				LRCN			
Datasets	Events	RGB	Depth	Infrared	Events	RGB	Depth	Infrared
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
Driver Drowsiness	63.5	52.7	55.6	53.2	55.6	54.2	55.4	55.7
Driver Gaze-Zone	77.1	79.0	72.0	78.5	91.7	69.5	92.7	91.2
Driver Hand-Gesture	63.6	90.9	92.2	91.1	73.7	88.8	93.3	76.7
	C3D				MiCT			
Datasets	Events	RGB	Depth	Infrared	Events	RGB	Depth	Infrared
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
Driver Drowsiness	73.5	81.6	62.1	68.6	60.6	56.2	57.7	61.5
Driver Gaze-Zone	97.0	97.0	90.1	98.4	75.2	90.9	90.5	92.7
Driver Hand-Gesture	89.0	94.5	96.1	95.2	63.3	84.4	82.4	85.6

TABLE VIII
COMPUTATION TIME ON THE NEUROIV DATASET. THE THIRD COLUMN SHOWS THE COMPUTATION TIME OF CORRESPONDING DATA FORMATS AND CONVOLUTIONAL ARCHITECTURES FOR DRIVER GAZE-ZONE DATASET

Datasets	Data	Computation Time(ms/frame)			
	Format	LSTM	LRCN	C3D	MiCT
Driver Gaze-Zone	DVS	1.76	1.58	9.31	0.44
	RGB	1.49	1.55	8.23	0.45
	Depth	2.01	1.54	4.48	0.48
	Infrared	2.01	0.98	8.33	0.50

changes, which can be further elaborated and discussed in future works.

VI. DISCUSSION

Neuromorphic vision is an emerging technology in the era of mature frame-based camera hardware and software. Comparisons are, in some terms, unfair since they are younger and smaller without caring out under the same maturity level. Thus, this work will act as the first database and baseline evaluations to fill the gap between neuromorphic engineering, intelligent vehicle and model computer vision. Since different communities focus on exploiting different advantages of the event-based paradigm, our work serves as a starting point to new researchers, giving a bird's-eye view to intelligent vehicle research community, and a new perspective to neuromorphic vision community.

There is no agreement on what is the best method to process events, notably because it depends on the applications. In this work we choose to convert event slice to event-frame mainly because almost all recent and successful deep learning algorithms are designed for frame-based video data to benefit from traditional processors. In order to take advantages of such techniques, and to provide a strong baseline of the dataset, asynchronous events are converted into synchronous event-frames with different encoding methods. Although the temporal aggregation of these methods needed to feed the network does, however, affect latency, a sliding mechanism could be adopted to split the asynchronous events and achieve an ultra-high frame-rate (e.g., 2ms sliding window equals to 500 fps).

Additionally, we notice that in CNN-LSTM network [62] there is a pre-trained Inception V3 [63] architecture learned on [7]. Even though it is well known that among different tasks feature learning is transferable, it is interesting to see that weights learned on standard RGB images have a successful transfer on encoded event-frames. As mentioned in [69], convolutional weights are sensitive to low-level features such as edges and boundaries in images, which is presented in encoded event-frames.

As a new database and baseline evaluations, there are many ways for researchers to work on the NeuroIV dataset. Firstly, to develop novel convolutional neural network is a straight way to improve the performance compared with our baseline results. Since we will release all of our dataset and models, we strongly encourage to train the network model from our raw data. Secondly, recently we witness large margin improvement by combining different modalities in action recognition (such as RGBD action recognition) as they are complementary in different aspects. The fusion of event frames with other modalities will be helpful for the improvement of the final performance. For example, the event frames are sensitive to the boundaries but with no depth information, while the depth images provide depth information for each event frame. The combination of the event frames and depth images deserves to be studied. Lastly, inspired by our experimental results that a simple encoding method can achieve comparable performance with depth images and infrared images, to discover a powerful encoding method will be worthy of study. A recent work from [36] shows that their reconstruction network is able to synthesize high framerate videos (5,000 frames per second) of high-speed phenomena (e.g. a bullet hitting an object) and is able to provide high dynamic range reconstructions in challenging lighting conditions. They demonstrate the effectiveness of their reconstructions as an intermediate representation for event data. They show that off-the-shelf computer vision algorithms can be applied to the reconstructions for tasks such as object classification and visual-inertial odometry and that the reconstruction images consistently outperforms algorithms that are specifically designed for event data. We therefore would highly recommend to try the work [36] with our NeuroIV dataset.

VII. CONCLUSION

It is acceptable that intelligent vehicles need to handle complex scenarios and, more importantly corner cases in which intelligent vehicle will maneuver. Exploring alternative approaches of neuromorphic vision sensor instead of developing algorithms of conventional cameras is of great value that can not only provide a complementary sensor to handle corner cases (as shown in Fig. 1) but also improve the robustness and accuracy of the performance in complex scenarios. In this work, we build the first-ever database, NeuroIV, and provides some baseline evaluations that bridges the gap between neuromorphic engineering and intelligent vehicle research. The NeuroIV introduces new ways to sense and perceive the environment that brings new revolution of vision-based perception system in intelligent vehicle. It will serve as a standardized and open-source platform on which new neuromorphic vision based methods can be developed and evaluated.

ACKNOWLEDGMENT

Guang Chen proposed the idea, defined the framework, collected the dataset, performed the experiments and wrote the paper. Fa Wang, Weijun Li, and Lin Hong collected the data and performed the experiments. Jieneng Chen, Zhenyan Zhang, and Yiwen Lu participated in

the collection of the data. Guang Chen, Jörg Conradt, and Alois Knoll supervised this work.

References

[1] G. Velez and O. Otaegui, "Embedding vision-based advanced driver assistance systems: A survey," *IET Intell. Transp. Syst.*, vol. 11, no. 3, pp. 103–112, Apr. 2017.

[2] Y. Liu *et al.*, "Motor-imagery-based teleoperation of a dual-arm robot performing manipulation tasks," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 3, pp. 414–424, Sep. 2019.

[3] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, and Q. Dai, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 284–295, Jan. 2018.

[4] G. Chen *et al.*, "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Jul. 17, 2020, doi: 10.1109/TSMC.2020.3005231.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[6] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[8] E. D. Dickmanns and V. Graefe, "Dynamic monocular machine vision," *Mach. Vis. Appl.*, vol. 1, no. 4, pp. 223–240, Dec. 1988.

[9] C. Urmson *et al.*, "Autonomous driving in urban environments: Boss and the urban challenge," *J. Field Robot.*, vol. 25, no. 8, p. 425–466, 2008.

[10] C. Brandli, R. Berner, M. Yang, S. Liu, and T. Delbruck, "A 240 × 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, Oct. 2014.

[11] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrbein, 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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," *Frontiers Neurosci.*, vol. 13, p. 73, Feb. 2019.

[16] 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/CVF Conf. Comput. Vis. Pattern Recognit.* Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 1731–1740.

[17] A. Z. Zhu, D. Thakur, T. Ozaslan, 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, Jul. 2018.

[18] A. I. Maqueda, A. Loquercio, G. Gallego, N. Garcia, 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.* Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 5419–5427.

[19] P. Gershon *et al.*, "Distracted driving, visual inattention, and crash risk among teenage drivers," *Amer. J. Preventive Med.*, vol. 56, no. 4, pp. 494–500, Apr. 2019.

[20] M. Knapik and B. Cyganek, "Driver's fatigue recognition based on yawn detection in thermal images," *Neurocomputing*, vol. 338, pp. 274–292, Apr. 2019.

[21] Y.-K. Wang, T.-P. Jung, and C.-T. Lin, "EEG-based attention tracking during distracted driving," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 6, pp. 1085–1094, Nov. 2015.

[22] Z. Li, S. Bao, I. V. Kolmanovsky, and X. Yin, "Visual-manual distraction detection using driving performance indicators with naturalistic driving data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 8, pp. 2528–2535, Aug. 2018.

[23] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 596–614, Jun. 2011.

[24] G. Wiesmann, S. Schraml, M. Litzenberger, A. N. Belbachir, M. Hofstatter, and C. Bartolozzi, "Event-driven embodied system for feature extraction and object recognition in robotic applications," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 76–82.

[25] E. Piatkowska, A. N. Belbachir, S. Schraml, and M. Gelautz, "Spatiotemporal multiple persons tracking using dynamic vision sensor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 35–40.

[26] G. Chen *et al.*, "Neuromorphic vision based multivehicle detection and tracking for intelligent transportation system," *J. Adv. Transp.*, vol. 2018, pp. 1–13, Dec. 2018.

[27] M. Litzenberger *et al.*, "Estimation of vehicle speed based on asynchronous data from a silicon retina optical sensor," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2006, pp. 653–658.

[28] P. Rogister, R. Benosman, S.-H. Ieng, P. Lichtsteiner, and T. Delbruck, "Asynchronous event-based binocular stereo matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 2, pp. 347–353, Feb. 2012.

[29] S. Schraml, A. N. Belbachir, and H. Bischof, "Event-driven stereo matching for real-time 3D panoramic vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 466–474.

[30] N. Matsuda, O. Cossairt, and M. Gupta, "MC3D: Motion contrast 3D scanning," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, Apr. 2015, pp. 1–10.

[31] B. Kueng, E. Mueggler, G. Gallego, and D. Scaramuzza, "Low-latency visual odometry using event-based feature tracks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 16–23.

[32] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 994–1001, Apr. 2018.

[33] 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.

[34] J. Binas, D. Neil, S.-C. Liu, and T. Delbrück, "DDD17: End-to-end DAVIS driving dataset," in *Proc. Workshop Mach. Learn. Auto. Vehicles (ICML)*, vol. abs/1711.01458, 2017, pp. 1–9.

[35] A. Zihao Zhu, Y. Chen, and K. Daniilidis, "Realtime time synchronized event-based stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 433–447.

[36] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 31, 2019, doi: 10.1109/TPAMI.2019.2963386.

[37] M. Ramzan, H. U. Khan, S. M. Awan, A. Ismail, M. Ilyas, and A. Mahmood, "A survey on state-of-the-art drowsiness detection techniques," *IEEE Access*, vol. 7, pp. 61904–61919, 2019.

[38] B. Mandal, L. Li, G. S. Wang, and J. Lin, "Towards detection of bus driver fatigue based on robust visual analysis of eye state," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 545–557, Mar. 2017.

[39] R. Oyini Mbouna, S. G. Kong, and M.-G. Chun, "Visual analysis of eye state and head pose for driver alertness monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1462–1469, Sep. 2013.

[40] W. Zhang and J. Su, "Driver yawning detection based on long short term memory networks," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–5.

[41] J. Lyu, Z. Yuan, and D. Chen, "Long-term multi-granularity deep framework for driver drowsiness detection," 2018, *arXiv:1801.02325*. [Online]. Available: http://arxiv.org/abs/1801.02325

[42] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Proc. Asian Conf. Comput. Vis. Workshop Driver Drowsiness Detection Video*, Taipei, Taiwan, Nov. 2016.

[43] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim, "Real-time gaze estimator based on Driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 254–267, Mar. 2011.

[44] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2074–2083.

[45] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.

[46] M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara, "From depth data to head pose estimation: A siamese approach," 2017, *arXiv:1703.03624*. [Online]. Available: http://arxiv.org/abs/1703.03624

[47] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[48] T. Baltrusaitis, P. Robinson, and L. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2610–2617.

[49] G. Borghi, R. Gasparini, R. Vezzani, and R. Cucchiara, "Embedded recurrent network for head pose estimation in car," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1503–1508.

[50] C. Ahlstrom, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 965–973, Jun. 2013.

[51] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 3, pp. 254–265, Sep. 2018.

[52] E. Tsironi, P. Barros, C. Weber, and S. Wermter, "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition," *Neurocomputing*, vol. 268, pp. 76–86, Dec. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231217307555

[53] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 409–419.

[54] H. Wang, P. Wang, Z. Song, and W. Li, "Large-scale multimodal gesture recognition using heterogeneous networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*. Venice, Italy: IEEE, Oct. 2017, pp. 3129–3137. [Online]. Available: http://ieeexplore.ieee.org/document/8265581/

[55] J. H. Lee *et al.*, "Real-time gesture interface based on event-driven processing from stereo silicon retinas," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2250–2263, Dec. 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6774446/

[56] E. Mueggler, C. Bartolozzi, and D. Scaramuzza, "Fast event-based corner detection," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2017, pp. 1–12.

[57] A. N. Burkitt, "A review of the Integrate-and-fire neuron model: I. Homogeneous synaptic input," *Biol. Cybern.*, vol. 95, no. 1, pp. 1–19, Jul. 2006.

[58] G. Chen *et al.*, "Multi-cue event information fusion for pedestrian detection with neuromorphic vision sensors," *Frontiers Neurorobotics*, vol. 13, p. 10, Apr. 2019.

[59] N. F. Y. Chen, "Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion," 2017, *arXiv:1709.09323*. [Online]. Available: http://arxiv.org/abs/1709.09323

[60] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[61] P. Bao, A. I. Maqueda, C. R. del-Blanco, and N. García, "Tiny hand gesture recognition without localization via a deep convolutional network," *IEEE Trans. Consum. Electron.*, vol. 63, no. 3, pp. 251–257, Aug. 2017.

[62] G. Thung and H. Jiang, "A torch library for action recognition and detection using CNNs and LSTMs," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2016.

[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[64] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[65] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[67] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[68] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "MiCT: Mixed 3D/2D convolutional tube for human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[69] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*. Washington, DC, USA: IEEE Computer Society, Jun. 2014, pp. 512–519.

Guang Chen (Member, IEEE) received the B.S. and M.Eng. degree in mechanical engineering from Hunan University, China, and the Ph.D. degree from the Faculty of Informatics, Technical University of Munich, Germany. He was a Research Scientist with fortiss GmbH, a research institute of the Technical University of Munich, from 2012 to 2016, and a Senior Researcher with the Chair of Robotics, Artificial Intelligence, and Real-Time Systems, Technical University of Munich, from 2016 to 2017. He is currently a Research Professor with Tongji University and a Senior Research Associate (Guest) with the Technical University of Munich. He is also leading the Intelligent Sensing, Perception and Computing Group, Tongji University. His research interests include computer vision, image processing and machine learning, and the bio-inspired vision with applications in robotics and autonomous vehicles. He was awarded the Program of Tongji Hundred Talent Research Professor 2018.

Fa Wang received the B.E. degree in vehicle engineering from Tongji University, Shanghai, China, where he is currently pursuing the master's degree in vehicle engineering with the Institute of Intelligent Vehicle. His research interests include intelligent vehicle perception systems based on computer vision and LiDARs.

Weijun Li received the B.E. degree in automotive engineering from Tongji University, Shanghai, China, where he is currently pursuing the M.Eng. degree in automotive engineering. His research interests include intelligent vehicle and computer vision.

Lin Hong received the B.E. degree in mechatronic engineering from the Shandong University of Science and Technology, Qingdao, China, where he is currently pursuing the M.Eng. degree in vehicle engineering. From 2018 to 2020, he was with the Institute of Intelligent Vehicles, Tongji University, Shanghai, China. His research interest includes safe driving, drowsiness driving detection, computer vision, and machine learning. His awards and honors include the National Scholarship for Postgraduates.

Jörg Conradt (Senior Member, IEEE) received the Diploma degree in computer engineering from TU Berlin, the M.S. degree in computer science and robotics from the University of Southern California, and the Ph.D. degree in physics/neuroscience from ETH Zurich. He is currently an Associate Professor with the School of Electrical Engineering and Computer Science, KTH, Stockholm, Sweden. Before joining KTH, he was a W1 Professor with the Technische Universität München, Germany. He was the Founding Director of the Elite Master Program NeuroEngineering, TUM. The research in his group focuses on neurocomputing systems and investigates key principles by which information processing in the brain works, and applies those to real-world interacting technical systems.

Jieneng Chen received the bachelor's degree in computer science from Tongji University, Shanghai, China.

Zhenyan Zhang received the bachelor's degree in computer science from Tongji University, Shanghai, China.

Yiwen Lu is currently pursuing the bachelor's degree in computer science with Tongji University.

Alois Knoll (Senior Member, IEEE) received the M.Sc. degree in electrical/communications engineering from the University of Stuttgart, Germany, in 1985, and the Ph.D. degree *(summa cum laude)* in computer science from the Technical University of Berlin, Germany, in 1988. He has been a Faculty Member with the Computer Science Department, TU Berlin, since 1993. He was with the University of Bielefeld, as a Full Professor and the Director of the Research Group Technical Informatics, since 2001. Since 2001, he has been a Professor with the Department of Informatics, TU München. He was also the Board of Directors with the Central Institute of Medical Technology, TUM (IMETUM). From 2004 to 2006, he was the Executive Director with the Institute of Computer Science, TUM. His research interests include cognitive, medical and sensor-based robotics, multi-agent systems, data fusion, adaptive systems, multimedia information retrieval, model-driven development of embedded systems with applications to automotive software and electric transportation, and simulation systems for robotics and traffic.