IEEE Open Journal of Vehicular Technology

Received 9 October 2023; accepted 16 October 2023. Date of publication 18 October 2023;
date of current version 7 November 2023. The review of this article was coordinated by Editor Shabnam Sodagari.

Digital Object Identifier 10.1109/OJVT.2023.3325656

Neuromorphic Driver Monitoring Systems: A Computationally Efficient Proof-of-Concept for Driver Distraction Detection

WASEEM SHARIFF [1,2], MEHDI SEFIDGAR DILMAGHANI [1], PAUL KIELTY [1], JOE LEMLEY [2], MUHAMMAD ALI FAROOQ [1], FAISAL KHAN [1], AND PETER CORCORAN [1] (Fellow, IEEE)

[1]Department of Electronic and Electrical Engineering, University of Galway, H91TK33 Galway, Ireland
[2]OCTO Sensing Team, Xperi Inc., H91V0TX Galway, Ireland

CORRESPONDING AUTHOR: WASEEM SHARIFF (e-mail: w.shariff1@universityofgalway.ie).

This work was supported by Irish Research Council to the employment-based PhD Scheme under Grant EBPPG/2022/17.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Xperi Research Ethics Committee.

ABSTRACT Driver Monitoring Systems (DMS) represent a promising approach for enhancing driver safety within vehicular technologies. This research explores the integration of neuromorphic event camera technology into DMS, offering faster and more localized detection of changes due to motion or lighting in an imaged scene. When applied to the observation of a human subject event camera provides a new level of sensing capabilities over conventional imaging systems. The study focuses on the application of DMS by incorporating the event cameras, augmented by submanifold sparse neural network models (SSNN) to reduce computational complexity. To validate the effectiveness of proposed machine learning pipeline built on event data we have opted the Driver Distraction as a critical use case. The SSNN model is trained on synthetic event data generated from the publicly available Drive&Act and Driver Monitoring Dataset (DMD) using a video-to-event conversion algorithm (V2E). The proposed approach yields comparable performance with state-of-the-art approaches, achieving an accuracy of 86.25% on the Drive&Act dataset and 80% on comprehensive DMD dataset while significantly reducing computational complexity. In addition, to demonstrate the generalization of our results the network is also evaluated using locally acquired event dataset gathered from a commercially available neuromorphic event sensor.

INDEX TERMS Distraction recognition, driver monitoring system driver monitoring system (DMS), event based vision, neuromorphic sensing, submanifold convolutions, computational complexity.

I. INTRODUCTION

In recent years, the integration of computer vision technologies into the automotive industry has gained momentum, driven by goals of improved driver safety and enhanced human-vehicle interactions [1], [2]. Within driver monitoring system, detecting distractions holds pivotal importance due to its role in traffic accidents [3]. Conventional camera sensors, including visible cameras, radar, and lidar, encounter a notable challenge arising from the temporal gaps between frames recorded by these sensors [22]. This "undersampling" can cause important information to be lost between frames [22], which is crucial for accurately detecting driver distraction.

Presently, the industry is directing its efforts towards advancing driver monitoring systems by leveraging neuromorphic vision technology [8]. This new approach takes ideas from neuromorphic event-based vision sensors to bring fresh perspectives to driver monitoring [6], [7]. These sensors operate differently from traditional cameras, capturing changes in brightness instead of complete frames at fixed intervals. They generate events or spikes when significant brightness changes occur, providing essential information such as pixel coordinates (location of the change), timestamps (corresponding timestamp of the change), and polarity (whether it was a positive change or negative change) [23].

While traditional approaches employ multi-sensor fusion and advanced algorithms to mitigate these challenges, event-based neuromorphic vision sensors offer a unique approach. They capture visual data as events triggered by significant brightness changes, sidestepping "undersampling" and motion blur challenges while achieving real-time detection of rapid luminance changes [22]. Moreover, event camera manufacturers provide users with a certain level of flexibility (Event-Bias setting) to customize the sensitivity of the event camera. In practical use cases these can be configured dynamically to handle variations in lighting conditions, and motion sensitivity. Practical examples include adaptation to low-light environments, high-contrast scenarios due to direct sunlight, or vehicular headlights. [24], [25], [26], [27]. This facilitates real-time detection and holds the potential to reshape sensor systems in the context of advanced optical sensing based vehicular technologies, marking a path beyond traditional algorithmic solutions.

In this context, event cameras, which selectively capture relevant changes in brightness, significantly reduce data throughput and computational requirements. Notably, in instances of motion absence, no unnecessary data is generated. This inherent sparsity in data stream translates into reduced processing power, memory, and bandwidth demands, resulting in potentially significant cost savings in terms of hardware infrastructure. Additionally, their ability to operate with low power consumption can lead to extended battery life in portable applications. While the initial investment in neuromorphic camera technology might be higher due to its specialized nature, the long-term cost benefits can outweigh the upfront expenditure, particularly in scenarios where computational efficiency and power conservation are paramount [23]. These advantages highlight the effectiveness of event cameras compared to traditional camera sensors. While event cameras offer significant advantages over traditional camera sensors, their sparse and asynchronous properties introduce distinctive challenges that require the innovation of novel processing techniques to adeptly analyze event data. At the same time they offer much greater temporal resolution than conventional camera sensors and individual events can be resolved at the microsecond level [6]. The core challenge is to sort the relevant event data for a particular sensing application from a much larger asynchronous event stream.

In this study, we focus on the development of submanifold sparse neural networks with event cameras to address the computational efficiency challenges. Inspired by the submanifold sparse convolutional neural network (SSCNN) and Asynet [13], [15], we propose a sparse-ResNet architecture as a binary classifier for efficient driver distraction prediction. This architecture optimally exploits the sparseness of event data, enhancing feature extraction while minimizing latency. By combining SSCNN with event camera sensing we significantly reduce the complexity of the computational analysis enabling efficient, low-power neuromorphic sensing techniques to be deployed in application fields such as driver monitoring and human activity sensing.

A key challenge in event camera research is the limitations of event-based datasets. To overcome this limitation, we employ an event simulator called v2e, to synthesize event-based datasets from visible video or image sequence data. By utilizing v2e, we create two synthetic event datasets: Drive&Act and Driver Monitoring Dataset (DMD) [11], [12]. These datasets allow for the evaluation and validation of our proposed approach in realistic driver distraction scenarios. Nine quantitative metrics, including accuracy, specificity, sensitivity, false positive rate, false negative rate, Matthews Correlation Coefficient, recall rate, F1 score and FLOPs(floating-point-operations) are utilized to analyze the performance of the classifiers. The results demonstrate the capabilities of the proposed submanifold sparse neural networks in efficiently detecting driver distraction. Finally the trained network is tested on locally acquired real-event dataset to show the generalization capabilities of the trained network.

Contributions:

1) Development of a sparse-ResNet architecture as a binary classifier to predict driver distraction, which can efficiently obtain features from a sequence of events and make use of the sparse signal for least computational processing.
2) Synthesis of two different RGB datasets to event datasets, Drive&Act and Driver Monitoring Dataset (DMD) to overcome data limitations and to introduce enough data diversity, using an event simulator called v2e, which allows for further training and testing of event algorithms.
3) Validation of low-power neuromorphic sensing through combining event-camera technology with SSCNN; a significant improvement in computational efficiency is demonstrated with an average of 1.4 GFLOPs.

II. RELATED WORK

A. CONVENTIONAL DRIVER DISTRACTION METHODS

Distracted driving raises the likelihood of serious accidents [1]. Since the rapid development of deep learning algorithms over the previous decade, distracted driving detection has motivated researchers' interest. In [17], the authors presented distraction detection based on kinematic motion models by fusing various state-space models to capture multiple driving motion patterns under ordinary driving conditions. Mustafa et al. [18], proposed E2DR - an ensemble technique to detect distraction in drivers. The E2DR variant with ResNet50 and VGG16 achieved 92% test accuracy. Authors in [19] proposed a driver distraction study on young-experienced drivers. There were two types of driving situations studied: manual driving conditions and partially automated driving conditions.

Kapotaksha et al. [16] proposed the detection and recognition of driver distraction using multi modal signals like visual, physiological and thermal groups of features with random forest and gradient boosting classifiers. Authors in [20], proposed an unsupervised deep learning algorithm that

applies RepMLP-Res50 (replaces some blocks in ResNet50 with ResMLP). This network was trained on a state-farm-distracted driver detection dataset and used 10%, 20%, 30%, 40%, and 50% of labelled data to fine-tune the network. In [21], the authors proposed a comprehensive review of the distraction related to mobile phone use. A systematic review of 37 papers was conducted using the PRISMA approach.

B. EVENT-BASED DRIVER MONITORING SYSTEM

In this section, some of the most recent event-based vision in context of driver monitoring systems is explored.

Event-based vision for driver monitoring has many advantages including high temporal resolution, reduced motion blur, low latency, reduced power consumption for the onboard sensor suite, and most importantly privacy [22], [23], [9], [6], [2]. The following are some of the recent studies on driver monitoring using event cameras. Authors in [6], [28], proposed a real-time face, eye tracking and blink detection using event cameras. Due to the limitation of the real face event dataset, the authors used the V2E simulator [10] to process events from the visible Helen dataset. Further, by exploiting the inherent advantages of vision sensors, Chen et al. [29] proposed event-based driver drowsiness detection using facial motion analysis. With the advantages of a high dynamic range and challenging environmental conditions like low illumination, authors in [30] proposed facial micro-expressions with an event camera. With the help of the signal produced by a high-speed event camera, the suggested approach shows how simple it is to understand the underlying emotions of faces that are being observed compared to conventional visible sensors. Authors in [31] proposed event-based near-eye gaze tracking beyond 10,000 Hz. A hybrid event-based eye-tracker which demonstrates a binocular prototype, functioning at a high framerate. Moreover, there are a few more studies with event cameras which can be efficiently used for driver monitoring systems [33], [34].

Recent studies have focused on leveraging event-based cameras, in more rigorous tasks. Berlincioni et al. [43] presented NEFER, a dataset for neuromorphic event-based facial expression recognition, highlighting the effectiveness of event-based approaches for analyzing subtle micro-expressions. Bulzomi et al. [44] introduced an end-to-end neuromorphic lip reading model, utilizing events captured by an event-based sensor for real-time embedded scenarios. Bissarinova et al. [45] released FES, the first large and varied dataset with face and facial landmarks annotations for event-based cameras, accompanied by models achieving high accuracy in predicting bounding boxes and landmarks. Further, Paul et al. [46], [47] proposed neuromorphic sensing techniques to analyze the entire facial region, detecting yawning behaviors that give a complimentary indicator of fatigue and drowsiness and similar approach to detect driver seatbelt. These recent studies showcase the increasing interest in utilizing event-based cameras for a wide range of computer vision applications.

To the best of the author's knowledge, there has been only one study focusing on event-based driver distraction recognition. Chu Yang et al. [35] suggested using EfficientNetB0 and LSTM cell to predict event-based driver distraction and action recognition. The v2e event simulator is used in the research to simulate the Drive&Act dataset while training on the simulated event dataset. Furthermore, study evaluates performance on a locally-acquired event dataset.

III. METHODOLOGY

In this section, a detailed overview of the proposed methodology including its event representation, and network architecture is described.

A. EVENT REPRESENTATION

Neuromorphic event cameras respond both independently and asynchronously to changes in illumination in the field of view. The event camera produces a series of variable data known as "Spike" or "Event" with each pixel-level change in the intensity [23]. When a sensor generates such an event, it acquires the logged brightness and continuously checks for a change from the brightness value it has previously saved. When a certain brightness threshold is exceeded, then the camera fires an event with the x, y location, the timestamp t, and the 1-bit polarity of the changes, indicating whether there was a positive change in brightness (1) or a negative change (0/−1). For one single event at a time interval ΔT a neuromorphic event camera produces the following

$$\{E_s\}_{s=1}^{N} = \{x_s, y_s, t_s, p_s\}_{k=s}^{N} \tag{1}$$

Driver distraction is a critical task for road safety, and it requires an understanding of both temporal and spatial information to predict distraction [35]. In this research, we used time-surface event representation, as it allows for the extraction of temporal information and pixel motion history directly from the event data. This can be achieved by grouping events in batches using a range of temporal bins, which enables the training of a neural network to learn spatiotemporal features. The advantage of using a time-surface representation for event data in submanifold sparse ResNet is that it eliminates the need for additional techniques like optical flow or 3D convolution, used in most detection networks to capture motion features explicitly [36]. This representation preserves the spatiotemporal sparseness of the event data, allowing for the least latency processing. Additionally, the submanifold sparse ResNet architecture is designed to handle the sparsity of the event data, resulting in a more efficient and effective classifier [39].

B. PROPOSED METHOD

In this research, we simulate two different datasets Drive&Act (side view - body focused) [11] and DMD (driver Monitoring dataset and this dataset has a front view of the driver - face focused) [12] from visible to event streams using V2E.

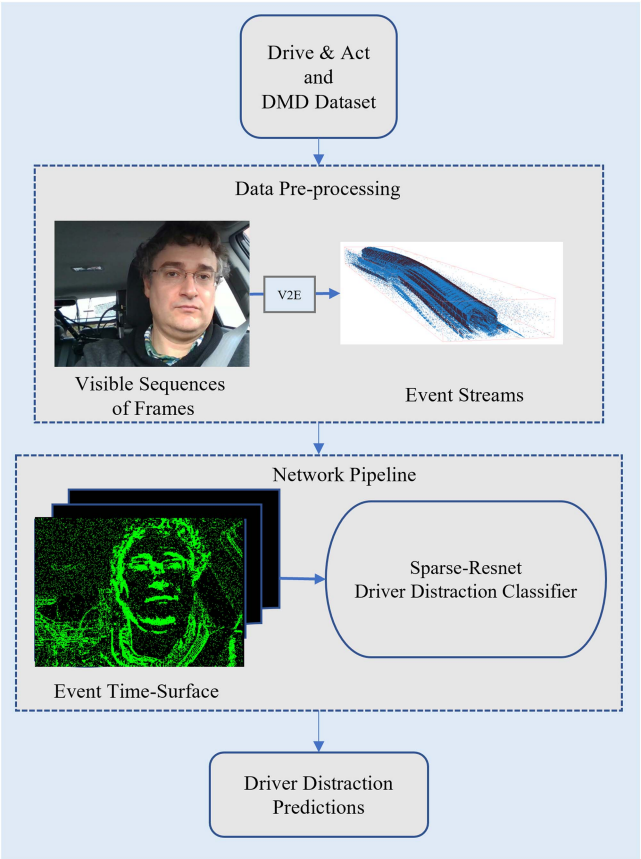

FIGURE 1. Overall pipeline of the proposed approach for event based driver distraction recognition.

Further, we train the simulated dataset separately on sparse-ResNet [15] architecture to further predict distraction. The same is depicted in the pipeline for event-based driver distraction in Fig. 1.

1) SUBMANIFOLD SPARSE CONVOLUTION NEURAL NETWORK (SSC)

The proposed approach utilizes the spatial sparsity of event data through Submanifold Sparse Convolutions (SSC), which significantly reduces computational requirements by focusing solely on activated sites. Unlike conventional convolutions, SSCs disregard inputs in the receptive field of the convolution that have zero values, calculating the convolution only at sites with non-zero feature vectors [13].

Fig. 2 presents the submanifold sparse ResNet architecture specifically designed to effectively process event data for predicting driver distraction. The figure consists of four parts, each representing a different step in the event processing through the network. In Fig. 2(a), an example input frame from a camera feed is shown. Fig. 2(b) illustrates the accumulation of events in a time-surface representation, showcasing the activated event sites (resulting from V2E tool which used to simulate RGB to events conversion). This step is crucial for converting event data into a format suitable for neural network processing. Fig. 2(c) depicts how the submanifold sparse ResNet architecture focuses only on the activated events, a key aspect that enables efficient event data processing. The architecture updates only the non-zero elements of the convolutional filters, significantly reducing computational overhead and memory requirements.

Moreover, in Fig. 2(d), we can observe a comparison between the updating mechanisms of conventional convolutions and the sparse-updating approach employed by the submanifold sparse ResNet architecture (depicted in Fig. 2(c)). This comparison brings attention to the disparities between these two approaches, specifically emphasizing the traditional convolutional updating method commonly used in conventional neural networks. Moreover, this comparison underscores the unique advantages offered by the submanifold sparse ResNet architecture in effectively managing the sparsity of event data. The use of a time-surface representation for event data in the architecture enables efficient feature extraction from event sequences. This representation converts the event sequence into a 2D spatiotemporal signal that can be processed by convolutional neural networks (CNNs) [39]. It preserves the spatiotemporal sparsity of the event data, allowing for minimal latency processing. Moreover, the design of the submanifold sparse ResNet architecture specifically addresses the sparsity of event data, resulting in a more efficient and effective classifier.

As mentioned before in SSCs, convolution operations are performed only at sites where there are non-zero feature vectors, referred to as active sites. These active sites are determined based on the presence of events. SSCs use a rulebook to establish correspondences between input and output sites. This rulebook dictates how information from active input sites is propagated to output sites, ensuring that computations are focused only on relevant areas. The initial Rulebook, proposed by [15], was later modified for event-based data analysis by [13], [14]. We have incorporated this adapted Rulebook to enhance our approach to driver distraction detection. The core equations governing SSCs in event-based data adapted from [13] are as follows:

- *Sparse Convolution Operation:*

$$y^t_{n+1}(u,c) = b_n(c) + \sum_{c_0}\sum_{k\in K_n}\sum_{(i,u)\in R_{t,k}} W_n(k,c_0,c)\,y^t_n(i,c_0),$$
$$\text{for } u \in A_t \tag{2}$$

Equation (2), $y^t_{n+1}(u,c)$ represents the activation at a specific pixel u and feature channel c in layer n at time $t+1$ and activated sites A_t. The term, $bn(c)$ is the bias term, which adds a constant value to the activation. $\sum_{c0}$ sums over all feature channels, ensuring contributions from different features are combined. $\sum_{k\in K_n}$ iterates over convolution kernel indices, allowing different kernels to be applied. $\sum_{(i,u)\in R_t,k}$ sums over the rulebook entries, where each entry (i,u) connects an active input

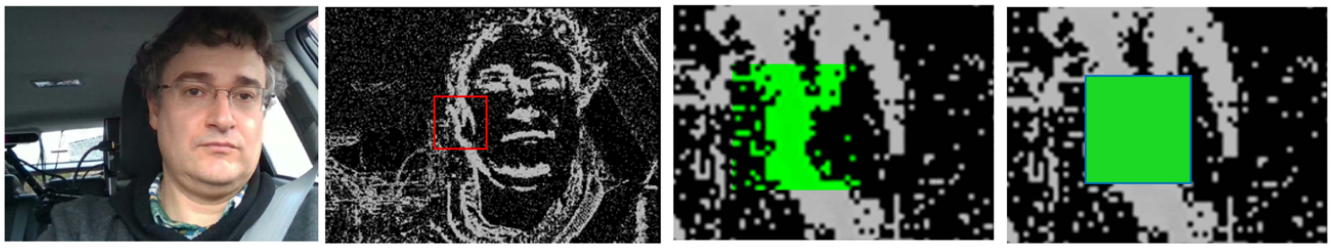

FIGURE 2. Proposed submanifold sparse approach for processing event data to detect driver distraction. The figure includes four components demonstrating different stages in the event processing pipeline. In panel, (a) an example input frame is presented, (b) shows how events are accumulated in a time-surface representation and activate non-zero event sites, (c) demonstrates how the submanifold sparse ResNet architecture handles sparse-updating, which enables efficient processing of event data by only updating non-zero elements of the convolutional filters, (d) compares the proposed sparse-updating approach to the conventional convolutional updating approach used in traditional neural networks.

site i to an output site u. The rulebook (R) preserves the connections between input and output sites, defining the relationships between input site i and output site j within R_t, k. A weighted summation is performed on the same output site $j = u$ to generate the integrated output. $W_n(k, c_0, c)$ represents the weight associated with a particular kernel index k and input/output feature channel pair (c_0, c). $y_t^n(i, c_0)$ is the activation at layer n for input site i and feature channel c_0 [13].

- *Activation Function:*

$$y_{t+1}^n(u, c_0) = \sigma(y_t^{n+1}(u, c_0)) \tag{3}$$

After the sparse convolution operation, the resulting values $y_{n+1}^t(u, c)$ (3) are passed through an activation function σ to introduce non-linearity into the network [13].

- *Update Rule for Active Sites:*

$$\Delta_n(u, c) = \sum_{k \in K_{n-1}} \sum_{(i,u) \in R_{k,n}} \sum_{c_0} W_{n-1}(k, c_0, c) * \$y_{n-1}^t(i, c_0) - y_{n-1}^{t-1}(i, c_0)) \tag{4}$$

From (4), $\Delta n(u, c)$ calculates the change in activation at active locations u in layer n due to a single event. $\sum_{k \in K^{n-1}}$ iterates over the convolution kernel indices from the previous layer, allowing the network to capture different spatial patterns. $\sum_{(i,u) \in R_k, n}$ sums over the rulebook entries specific to layer n and kernel index k, connecting input site i to output site u. $W_{n-1}(k, c_0, c)$ represents the weight associated with a particular kernel index k and input/output feature channel pair (c_0, c). $y_{t-1}^n(i, c_0)$ is the activation at layer n for input site i and feature channel c_0 at the previous time step. $y_{t-1}^n(i, c_0)$ reflects the previous state of active sites, allowing the network to efficiently update only the relevant locations in response to new events [13].

2) SPARSE-RESNET

SparseResNet is a specialized implementation of the Pre-activated Residual Network (ResNet) architecture, designed to address the challenge of vanishing gradients in deep neural networks, particularly for sparse data like event-based vision data from neuromorphic sensors. Traditional deep networks suffer from vanishing gradients as they get deeper, making training difficult [52]. ResNet introduces residual blocks with shortcut connections to enable the flow of gradients, addressing this issue effectively.

In this research we utilize the sparse ResNet architecture, introduced in [15], which is specifically designed to leverage the inherent spatial sparsity of data. Moreover, this research employed the ResNet basic "block" consists of two submanifold convolutional layers with batch normalization and ReLU activations, forming the fundamental building unit of ResNet architectures (basic block takes-input channels, number of repetitions and stride). The number of repetitions in each layer controls how many times the basic block is applied sequentially. Fig. 3(a) shows a single resnet basic block, followed by figure 3(b) which shows how multiple basic blocks are grouoped to predict distraction. The proposed network consists of five blocks with varying numbers of channels: 16, 32, 64, 128, and 256 as shown in figure 3(b). The input to the ResNet block maintains a resolution of 640 × 480 with 2 time-surface bins. Following the sparse ResNet block, the network undergoes an fully connected layer to generate predictions. In Pre-activated ResNet, batch normalization and ReLU activations are applied before the convolution operation, enhancing stability during training and improving overall performance (Fig 3(a)). Batch normalization is applied during training to normalize input data within each mini-batch, accelerating training and enhancing stability. The ReLU activation function introduces non-linearity, enabling the model to learn complex patterns.

IV. EXPERIMENTATION

A. DATASET

The detailed descriptions of the two datasets used for training the network is provided in Sections IV-A1 and IV-A2. To evaluate the performance, a 3-fold validation approach was employed on the dataset. Table 1 presents the breakdown of the three-fold splits for both the Drive&Act and DMD datasets. The Drive&Act dataset consists of data from 15 subjects, while the DMD dataset includes data from 20 subjects. The table highlights the specific splits utilized for

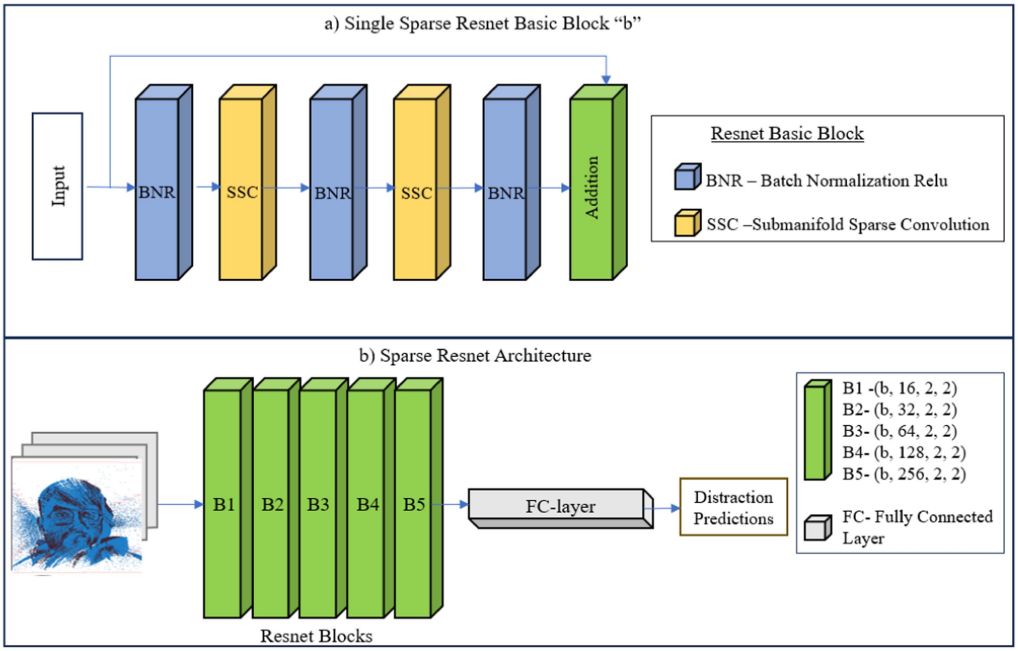

FIGURE 3. Submanifold Sparse Convolution ResNet: a) showcases a single ResNet block and b) presents the complete architecture including six basic blocks.

TABLE 1. Simulated Drive& Act and DMD Event Dataset 3 Different Splits

	Drive&Act (15 subjects)			DMD Dataset (20 subjects)		
	Train	Val	Test	Train	Val	Test
Split 0	Remaining	14, 15	5, 11, 13	Remaining	4, 11, 15	12, 14
Split 1	Remaining	3, 8	9, 10, 12	Remaining	5, 9, 16	10, 20
Split 2	Remaining	1, 2	4, 6, 7	Remaining	2, 6, 7	3, 8

training, validation, and testing in the three-fold training process.

1) DRIVE&ACT DATASET

Drive&Act dataset [11] is a domain-specific dataset that was specifically designed for fine-grained driver action recognition. The dataset consists of more than 9.6 million frames from 15 subjects. To generate this dataset, videos of subjects driving were recorded, and for each action, on average, a 3-second video was cropped and converted to event streams. Specifically, for each subject, there were 40 sequences of two different videos were trimmed. Half of the samples in this dataset consisted of various distractions, such as using a mobile phone, adjusting the radio, talking on the phone, reading a newspaper, fetching an object, eating, drinking, and/or using a laptop while driving. The other half of the sequences consisted of focused driving. In total, 560 samples were derived from the 15 subjects, and these samples were then converted to event streams using the v2e tool [10]. However, for subjects 9 and 10, due to the short length of their videos, only 40 sequences were extracted from each of them.

2) DRIVER MONITORING DATASET (DMD)

The Driver Monitoring dataset [12] is a comprehensive dataset that is widely used in the field of driver monitoring. This dataset is unique because it includes data collected in three different scenarios: a laboratory environment, a vehicle where subjects are acting, and real-world driving scenarios. The authors of [12] have made data from 20 subjects available for academic research. Each subject participated in three different driving scenarios, resulting in a total of 1680 video samples, each with an average duration of 3 seconds. The videos were cropped and simulated to event streams using v2e. The dataset includes a variety of driving scenarios, such as distracted driving where subjects are using mobile phones, drinking water, fixing their hair, talking on the phone, and fetching objects, as well as focused driving scenarios where the subjects are paying attention to the side mirrors and the road. The diversity of scenarios captured in this dataset is a significant advantage, as it allows the distraction detection model to be more robustly trained and evaluated using locally acquired event data. This dataset contains 50% of samples with distracted driving and 50% of samples with focused driving from all three scenarios. The fact that the dataset contains data acquired in a variety of circumstances makes it an ideal for this research. The trained

model can be effectively used to detect distracted driving, and its performance can be evaluated using locally acquired event data.

B. REAL-EVENT DATA ACQUISITION

In order to ensure that the trained network could generalize well to real-world scenarios, it was important to evaluate its performance on real event data. To this end, a new dataset was collected in a laboratory environment, where subjects were instructed to perform a variety of actions that might distract them while driving, such as texting, talking on the phone, fixing their hair, fetching objects, fixing the radio, drinking water, and so on. The data acquisition method used for this dataset was the same as that used for the DMD dataset, which consisted of face-focused recordings. A total of 20 subjects volunteered for the study, and for each subject, 10 sequences of distraction and 10 sequences of focused driving were recorded. The data was collected with informed consent and in compliance with ethical guidelines.

To record our real event driver distraction dataset, we utilized Prophesee's EVK3 Gen 3 camera [42] in Xperi Data Acquisition setup [38]. In Fig. 4 showcases the complete setup of our driving simulator for data acquisition. The driver simulator employed in this study is a state-of-the-art industry-standard system designed specifically for evaluating the latest advancements in Driver Monitoring Systems (DMS). This simulator is distinguished by its high fidelity, realism, and adherence to industry benchmarks, making it a paramount tool for replicating real-world driving scenarios and assessing the performance of state-of-the-art DMS technologies. For more information on the simulator please refer [38]. This high-speed vision sensor directly captures events instead of traditional frames. We collected a total of 400 sequences, encompassing diverse real-world driving scenarios. By incorporating this real event data during the evaluation of our trained network, we systematically measure robustness and effectiveness of proposed approach in addressing various distraction scenarios encountered in real-world driving situations. Fig. 5, which presents plots illustrating events represented during network learning. This figure encompasses samples from both the Simulated Event Dataset and locally acquired event datasets.

FIGURE 4. Real-events data acquisition setup using Prophesee EVK3.

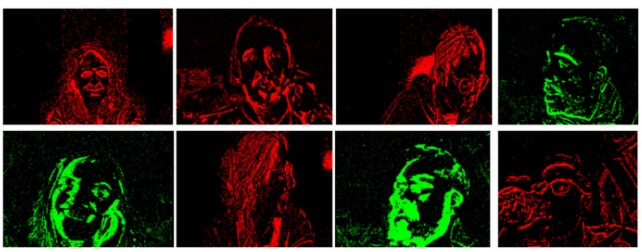

FIGURE 5. Visualised events generated from real and simulated event data.

C. IMPLEMENTATION SPECIFICATION

All the experiments were performed on Nvidia-RTX-2080Ti GPU with 64 GB RAM using the PyTorch framework. There are two aspects to the assessment of driver distraction detection. First, models were entirely trained on simulated event datasets and further, they were tested on unseen simulated data. Second, the model was tested using locally acquired event datasets and then assessed once again after fine-tuning the trained model with real event data. During the training process for the Drive&Act dataset of the research, the Adam optimizer was used with an initial learning rate of 1e-3 which is degraded with a factor of 0.1 every 50 epochs. For the training process of the DMD dataset which is face-focused, Adam optimizer was used with an initial learning rate of 1e-4 which was again degraded with a factor of 0.1 every 50 epochs. The training consisted of a batch size of 16 and was terminated after 65 and 85 epochs respectively. Since the trained network is a binary classifier, we used binary cross entropy as our loss function.

V. EVALUATION OF SIMULATED EVENT DATASET

To evaluate the efficacy of the proposed models, we adopted a evaluation strategy similar to that utilized in reference in [35].

TABLE 2. Quantitative Results on Simulated Event Datasets

Evaluation	Body-Focused	Face-Focused	Derivations
Sensitivity	0.8406	0.7804	TPR = TP / (TP + FN)
Specificity	0.8850	0.8226	SPC = TN / (FP + TN)
Precision	0.8910	0.8350	PPV = TP / (TP + FP)
Negative Predictive Value	0.7650	0.7810	NPV = TN / (TN + FN)
False Positive Rate	0.1150	0.1774	FPR = FP / (FP + TN)
False Discovery Rate	0.1090	0.1650	FDR = FP / (FP + TP)
False Negative Rate	0.1594	0.2196	FNR = FN / (FN + TP)
Accuracy	0.8615	0.8005	ACC = (TP + TN) / (P + N)
F1 Score	0.8650	0.8068	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.7245	0.6015	MCC = TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*TN+FN))

This 3-fold splitting mitigates the risk of overfitting to a specific set of data while offering a more reliable estimate of the model's performance across diverse scenarios and ensuring the robustness and reliability of the research findings. Similar to [35], split 0 was selected to evaluate the body-focused method, and the performance was compared against the state-of-the-art (SoA) approaches (Drive&Act dataset). On the other hand, to the authors' knowledge, this is the first instance of benchmarking accuracy and precision for the face-focused approach (DMD-dataset) using an event dataset. In this study, the performance of the proposed method was compared with an existing method, LSTM [35], which was considered as SoA. However, to ensure a fair comparison, it is crucial to not only evaluate the performance but also consider the computational complexity of both methods. Regrettably, the recent studies lack specific information regarding the number of parameters and FLOPs (floating-point operations) required for their approach. Hence, we implemented an approximation of these values based on the limited information provided in the referenced article [35].

This research also uses various quantitative metrics as evaluation indicators to assess the performance of trained classifiers. These metrics include accuracy, precision, sensitivity, specificity, false positive rate, false negative rate, and Matthews correlation coefficient. These indicators play a crucial role as a comprehensive set of evaluation metrics for measuring the effectiveness of the proposed driver distraction detection method. The details of these metrics are presented in Table 2 with corresponding results for face-focused and body-focused approaches.

Accuracy measures how often the model correctly identifies whether a driver is distracted or not. It provides an overall assessment of the classifier's ability to make correct predictions. *Precision*, on the other hand, evaluates the proportion of true positives (correctly identified distractions) out of all positive predictions. It helps assess the classifier's ability to minimize false positive errors by focusing on the precision of positive predictions. *Sensitivity*, also known as recall or true positive rate (TPR), measures the ability of the classifier to correctly identify positive instances (driver distractions) out of all actual positive cases. It indicates how well the classifier captures the positive examples. *Specificity* quantifies the ability of the classifier to correctly identify negative instances (non-distracted drivers) out of all actual negative cases. It represents the classifier's capacity to correctly exclude negative examples. *False positive rate (FPR)* computes the proportion of negative instances incorrectly classified as positive (non-distracted drivers classified as distracted). It highlights the rate of false alarms or Type I errors. *False negative rate (FNR)* calculates the proportion of positive instances incorrectly classified as negative (distracted drivers classified as non-distracted). It measures the rate of missed detections or Type II errors. *Matthews correlation coefficient (MCC)* combines information from true positives, true negatives, false positives, and false negatives to provide an overall measure of classifier performance. It takes into account imbalanced datasets and is particularly useful when the classes are not equally represented.

In addition to these metrics, the research also employed Flops (floating-point operations per second) as a measure to compare the computational complexity of the proposed method with that of a previously state-of-the-art method. Flops quantifies the number of arithmetic operations a computer can perform per second, serving as an indicator of the computational resources required to execute a specific method. It helps evaluate the efficiency and computational demands of different algorithms or approaches.

A. PERFORMANCE ON BODY-FOCUSED DRIVE&ACT DATASET

The evaluation of the submanifold sparse ResNet architecture commenced with testing its performance on an unseen Drive&Act simulated dataset. The results obtained from the driver distraction classification are presented in Table 3. In comparison to the state-of-the-art LSTM model, the sparse ResNet architecture showcased competitive performance. While the LSTM achieved an accuracy of 89.23%, the sparse ResNet achieved a comparable accuracy of 86.25%. Notably, it also demonstrated the highest precision of 0.89. It is important to emphasize that the proposed network was trained from scratch, devoid of any pre-trained network weights for the classification task. These findings underscore the suitability

TABLE 3. Performance Results on Simulated Drive&act and DMD Event Dataset: Evaluating the Effectiveness of the Proposed Sparse-ResNet Architecture (approximate Values*)**

Model	Pretrained	Dataset	Accuracy	Precision	Parameters	GFLOPs
LSTM [35]	ImageNet	Drive& Act	**89.23%**	**0.92**	6.3M*	478.47*
Sparse-ResNet (Proposed)	From-Scratch	Drive& Act	86.25%	0.89	**311K**	**1.39**
Sparse-ResNet (Proposed)	From-Scratch	DMD	**80.05%**	**0.83**	**311K**	**1.42**

of the submanifold sparse ResNet architecture for processing event data, as it effectively extracts features and produces accurate predictions.

Additionally, Table 2 provides a comprehensive evaluation of body-focused model. These metrics offer an intricate and multifaceted assessment of the body-focused techniques and their performance across the simulated event datasets. Specifically, the sensitivity metric, measuring the ability to accurately detect true positive cases, achieved a score of 0.8406. The specificity metric, indicating the ability to correctly identify true negative cases, reached a value of 0.8850. The false positive rate, representing the proportion of falsely identified positive cases among actual negatives, was calculated at 0.1150. Similarly, the false discovery rate, which signifies the proportion of falsely identified positive cases among all positive predictions, amounted to 0.1090. On the other hand, the false negative rate, representing the proportion of actual positive cases that were incorrectly predicted as negative, was observed at 0.1594.

Furthermore, the F1 score, which harmonizes the balance between precision and recall, achieved a value of 0.8650. This metric is particularly important as it conveys the trade-off between making accurate positive predictions and capturing all actual positive cases. Lastly, the Matthews Correlation Coefficient (MCC), which takes into account true and false positives and negatives to provide a comprehensive measure of classification performance, achieved a value of 0.7245. Collectively, these metrics offer a comprehensive and nuanced understanding of the efficacy of the body-focused methods and their proficiency in handling the simulated event datasets, shedding light on different aspects of their performance and highlighting their strengths and areas for potential improvement.

B. PERFORMANCE ON FACE-FOCUSED DMD DATASET

The DMD dataset is a robust simulated dataset comprising over 1680 samples that underwent a 3-fold validation criteria. It encompasses a diverse range of both real-world and laboratory driving scenarios, adding to its reliability. The performance evaluation of the trained sparse ResNet on this unseen dataset is presented in Table 3. The results indicate that the trained network achieved an average accuracy of 80.05% and a precision of 0.83 on the event-simulated DMD dataset.

In parallel, Table 2 presents an all-encompassing evaluation of face-focused methodologies on Simulated Event Datasets. The array of metrics within the table provides a comprehensive perspective on the performance of these face-focused techniques across the simulated event datasets. To delve into specifics, the sensitivity metric, gauging the capability to accurately detect true positive instances, achieved a score of 0.7804. Simultaneously, the specificity metric, reflecting the capacity to correctly identify true negative instances, reached a level of 0.8226. Further analysis indicates that the precision of the face-focused approaches attained a value of 0.8350, while the negative predictive value, representing the ability to correctly predict true negatives, reached a value of 0.7650. Additionally, the false positive rate, signifying the ratio of incorrectly predicted positive cases among actual negatives, was computed at 0.1744. Similarly, the false discovery rate, indicating the ratio of falsely predicted positive cases among all positive predictions, coincided at 0.1650. Notably, the false negative rate, denoting the proportion of actual positive cases incorrectly classified as negative, amounted to 0.2196. The F1 score, which balances precision and recall, culminated at a value of 0.8068. Not to be overlooked, the Matthews Correlation Coefficient (MCC), providing a comprehensive assessment of classification performance by considering true and false positives and negatives, reached a value of 0.6015. These metrics collectively offer an intricate and comprehensive evaluation of the face-focused methodologies, shedding light on diverse aspects of their performance across the simulated event datasets. Additionally, the pretrained weights of this network were fine-tuned on real event data to assess its performance in that domain as well.

C. COMPUTATIONAL COMPLEXITY AND ROBUSTNESS

The FLOPs of body-focused and face-focused networks have been studied and analyzed. The SSC library includes a namespace variable to calculate a network's FLOPs. Each time SparseConvNet does a sparse convolution forward operation with the specified MAC step count, the FLOPs are updated. To calculate the average FLOPs per forward pass for each accumulated time surface, the total number of test cases is divided, and the receptive fields marked as zero are excluded.

Due to resource constraints, conducting a direct performance comparison between our proposed model and other state-of-the-art models was not feasible. However, we estimated the GFLOPs for a comparable LSTM network based on information from a referenced paper [35]. Table 3 shows estimated GFLOPs (Giga Flops) for the comparable LSTM network were approximately 478, taking into account assumptions made about event representation and available details. On the other hand, during the training of the sparse-resnet model, we measured an average of 1.4 billion FLOPs (1.4 GFLOPs) per accumulated event representation.

TABLE 4. Existing Lightweight Networks in Terms of GFLOPS for Video Clip Based Processing (RGB)

Model	Modality	GFLOPs
C3D [48]	RGB	38.55
3D Student Network [51]	RGB	37.20
I3D [50]	RGB	27.90
P3D ResNet [49]	RGB	18.67
Proposed	Event	**1.42**

To gain a broader understanding and context of the current landscape of lightweight networks specifically focused on distraction classification, we examined recent developments in neural networks, as summarized in Table 4. This table presents the GFLOPs data associated with state-of-the-art models in this field. While it is important to note that Table 4 pertains to networks designed for a fine-grained classification task with more than 2 classes, it provides a valuable reference point for comparing computational complexities based on GFLOPs. By considering the GFLOPs values of the recent state-of-the-art networks in, we can observe that they tend to have significantly higher computational requirements compared to proposed event-based model. This finding suggests that these state-of-the-art networks may be less suitable for deployment in resource-constrained environments. It highlights the potential advantage of our proposed model, which offers a more lightweight alternative in terms of computational complexity.

From Table 3, when comparing the proposed Sparse Resnet model with the state-of-the-art LSTM model [35], some significant differences emerge. The LSTM model had 6.3 million parameters, achieved an accuracy of 89.23%, and operated at a GFLOPs rate of 478.47*. In contrast, our proposed Sparse Resnet model trained on the Drive&Act dataset consisted of only 311 thousand parameters, achieved an accuracy of 86.25%, and operated at a GFLOPs rate of 1.39. Similarly, when trained on the DMD dataset with the same 311 K parameters, our Sparse Resnet achieved an accuracy of 80.05% with 1.42 GFLOPs. Notably, the GFLOPs for the LSTM network were over two orders of magnitude higher than those of our proposed Sparse Resnet model, which emphasizes the computational efficiency and effectiveness of our model compared to the state-of-the-art approaches. This contribution is significant as it showcases the effectiveness of sparse networks in processing event-based data, reducing computational requirements while preserving high accuracy.

Factors: The lower computational complexity of the proposed sparse-ResNet architecture compared to SoA can be attributed to several factors. Firstly, by leveraging the sparsity, the sparse-ResNet architecture can significantly reduce the number of computations required during network training and inference. In contrast, conventional neural networks are designed to model sequential data and typically operate on dense input data. They require processing each element in the sequence, leading to higher computational requirements. Furthermore, the submanifold sparse convolution technique employed in the sparse-ResNet architecture emphasizes only the activated sites, further reducing computation and enhancing feature extraction while minimizing latency. This technique allows for selective processing of relevant event data, contributing to the overall computational efficiency of the network.

TABLE 5. Event Dataset Testing on Simulated DMD Pre-Trained Weights

Fine-Tune	Accuracy	Precision
initial-test	48%	0.52
20%	67%	0.68
50%	87%	0.85

VI. PERFORMANCE ON REAL-EVENT DATASET

Fine-tuning the Sparse-Resnet DMD model on real event data significantly enhances its performance and generalization capability for detecting driver distraction in diverse scenarios. The model's initial training on simulated data provides a foundation, but fine-tuning on the locally acquired event dataset exposes the network to real-world complexities and variations that were absent in the simulated training data. Moreover, by fine-tuning, the model adapts and learns from the specific features and patterns present in real event data, enabling it to better understand and distinguish different types of driver distractions in diverse scenarios.

In the experiments conducted on the locally acquired event dataset, the performance of the network was evaluated through three different stages. Initially, the pretrained weights of the Sparse-Resnet-DMD model were tested on the dataset in the first experiment. Table 5 presents the testing performances achieved on the locally acquired event dataset. The results obtained from this initial test showed a performance of 48% accuracy and a precision of 0.52. In the second and third experiments, the network was fine-tuned using a portion of the locally acquired event dataset. Specifically, in the second experiment, the network was trained on the first 20% of the data. The fine-tuning process improved the network's performance, resulting in an accuracy of 67% and a precision of 0.68. In the third experiment, the network was further fine-tuned on 50% of the locally acquired event data. This additional fine-tuning significantly improved the network's performance, with an accuracy of 87% and a precision of 0.85. During the training process, an initial learning rate of 1e-4 was used, and the network was trained for approximately 400 epochs.

The network's convergence during training, as evidenced by the improved performance on the real event dataset, further confirms the effectiveness of the Sparse-Resnet-DMD model. It successfully generalizes well on real event data despite being initially trained on simulated data. The ability of the model to perform consistently on both simulated and real event data highlights its robust and versatile nature, making it a promising tool for detecting driver distraction across diverse scenarios.

VII. LIMITATIONS

In the context of this work being an initial proof of concept for the use of event cameras (EC) and neural networks in driver distraction detection, it is important to acknowledge the associated limitations. These limitations serve as valuable areas for future improvement and exploration.

- *Challenges in providing diverse event representations:* Sparse submanifold ResNet may face challenges in effectively incorporating diverse event representations into the network. Event-based vision involves processing asynchronous and sparse data, where events are captured at irregular intervals and contain different levels of spatial and temporal information. Designing a network architecture that can handle such variability and effectively encode different event representations can be challenging.
- *Challenges in capturing temporal dependencies:* While sparse submanifold ResNet is designed to capture spatial relationships among events effectively, it may struggle to model long-term temporal dependencies. Distraction detection often requires understanding the temporal context of events and how they evolve over time. The limited temporal information available in event-based data can pose challenges in capturing complex temporal dependencies accurately in sparse submanifold networks.
- *Binary classification limitations:* The algorithm's binary classification approach may result in false positives, classifying non-distracting actions as distractions (such as fixing seat-belt or adjusting seat). Exploring nuanced classification approaches or incorporating context-aware features can help address this limitation in future iterations.
- *Limited robustness to occlusions and partial visibility:* The algorithm may struggle with occlusions or partial visibility of the driver's face or hands. Further research and algorithm refinements are needed to improve its handling of occlusions and partial visibility.

VIII. DISCUSSION

Simulating event data from regular videos proved to be valuable for training the model and further fine-tuning on a real event dataset collected locally. This highlights the potential of using simulated datasets as a cost-effective way to generate large amounts of training data, enabling the development of accurate and robust models. The ability to fine-tune the model on real event data enhances its performance and generalization capability in detecting driver distraction in diverse scenarios.

Further the proposed work also adopts the sparse-resnet architecture which is specifically designed to handle the sparse data. It incorporates the submanifold sparse convolution technique, which reduces computational requirements while improving feature extraction and minimizing latency. The combination of event cameras (EC) with the submanifold sparse convolutional neural network architecture offers computational efficiency, achieving more than two orders of magnitude improvement compared to LSTM-based approaches. Despite the higher accuracy of the current state-of-the-art model (89.23%), the Sparse-ResNet architecture outperforms it by offering fewer parameters (311 K vs. 6.3 M) and requiring lower computational resources (1.39 GFLOPs vs. 478.47 GFLOPs) on the Drive&Act dataset. This highlights the advantage of leveraging the unique properties of event cameras, such as low latency, high dynamic range, and reduced computational requirements, along with the sparse-ResNet architecture for efficient driver distraction detection.

IX. CONCLUSION

In this article, a proof-of-concept realization of a submanifold based neuromorphic event camera technology for driver distraction detection is evaluated. The research highlights the inherent advantages of event cameras, particularly their spatial sparsity, which significantly reduces computational complexity when combined with submanifold network models. With a computational complexity of approximately 1.4 GFLOPs, the proposed approach is computationally efficient alternative to equivalent models. Beyond their role in driver distraction detection, these event cameras have broader applications in enhancing driver monitoring systems for advanced vehicles [8]. The use of neuromorphic vision technology holds promise for improving human-vehicle interactions and safety, representing a transformative shift away from traditional algorithmic solutions. By leveraging the spatial sparsity of event data, this opens the door to real-time applications, bringing about significant changes in the realm of advanced optical sensing within the automotive industry.

A. FUTURE WORK

In future work, addressing the limitations addressed above will be a key focus. This includes exploring techniques to effectively incorporate diverse event representations and developing methods to capture long-term temporal dependencies. More sophisticated classification approaches and context-aware features will also be investigated to mitigate the limitations of binary classification and improve accuracy. Additionally, future work involves enabling asynchronous modification of events during inference, studying the relationship between driver activities and cognitive load, comparing computational complexity with state-of-the-art methods, and evaluating performance on embedded platforms. These platforms will serve as onboard sensor suites for real-time analysis with an evaluation of events per second. These efforts will contribute to further advancing the capabilities of the architecture and enhancing its applicability in driver distraction detection.

ACKNOWLEDGMENT

The authors acknowledge the significance of Xperi Inc. support in enabling the data acquisition setup, which allowed for the collection of reliable and comprehensive data. They also recognize the participants' crucial role in providing the necessary data that formed the foundation of their research. The combined efforts of Xperi and the study participants have significantly contributed to the advancement of knowledge in the field.

REFERENCES

[1] X. Wang, R. Xu, S. Zhang, Y. Zhuang, and Y. Wang, "Driver distraction detection based on vehicle dynamics using naturalistic driving data," in *Transp. Res. Part C, Emerg. Technol.*, vol. 136, 2022, Art. no. 103561.

[2] M. A. Farooq, W. Shariff, D. O'callaghan, A. Merla, and P. Corcoran, "On the role of thermal imaging in automotive applications: A critical review," *IEEE Access*, vol. 11, pp. 25152–25173, 2023, doi: 10.1109/ACCESS.2023.3255110.

[3] CDC, "Distracted driving." Apr. 26, 2022. Accessed: Mar. 9, 2023. [Online]. Available: https://www.cdc.gov/transportationsafety/distracted_driving/index.html

[4] S. Arun, K. Sundaraj, and M. Murugappan, "Driver inattention detection methods: A review," in *Proc. IEEE Conf. Sustain. Utilization Develop. Eng. Technol.*, 2012, pp. 1–6, doi: 10.1109/STUDENT.2012.6408351.

[5] A. Kashevnik, R. Shchedrin, C. Kaiser, and A. Stocker, "Driver distraction detection methods: A literature review and framework," *IEEE Access*, vol. 9, pp. 60063–60076, 2021, doi: 10.1109/ACCESS.2021.3073599.

[6] C. Ryan et al., "Real-time multi-task facial analytics with event cameras," *IEEE Access*, vol. 11, pp. 76964–76976, 2023, doi: 10.1109/ACCESS.2023.3297500.

[7] P. Kielty, M. S. Dilmaghani, W. Shariff, C. Ryan, J. Lemley, and P. Corcoran, "Neuromorphic driver monitoring systems: A proof-of-concept for yawn detection and seatbelt state detection using an event camera," *IEEE Access*, vol. 11, pp. 96363–96373, 2023, doi: 10.1109/ACCESS.2023.3312190.

[8] Prophesee, "Xperi develops world first neuromorphic in-cabin sensing (driver monitoring system)." Jan. 2023. [Online]. Available: https://www.prophesee.ai/xperi-develops-world-first-neuromorphic-in-cabin-sensing/

[9] W. Shariff, M. A. Farooq, J. Lemley, and P. Corcoran, "Event-based YOLO object detection: Proof of concept for forward perception system," in *Proc. SPIE*, vol. 12701, pp. 74–80, 2023.

[10] Y. Hu, S. C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1312–1321.

[11] M. Martin et al., "Drive&Act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2801–2810, doi: 10.1109/ICCV.2019.00289.

[12] J. D. Ortega et al., "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 387–405.

[13] N. Messikommer, D. Gehrig, A. Loquercio, and D. Scaramuzza, "Event-based asynchronous sparse convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 415–431.

[14] "Asynet- Rulebook." Apr. 7, 2021. Accessed: Oct. 06, 2023. [Online]. Available: https://github.com/uzh-rpg/rpg_asynet/tree/master/async_sparse_py/include/Rulebook.h

[15] B. Graham and L. v. d. Maaten, "Submanifold sparse convolutional networks," 2017, *arXiv:1706.01307*.

[16] K. Das et al., "Detection and recognition of driver distraction using multimodal signals," *ACM Trans. Interactive Intell. Syst.*, vol. 12, no. 4, pp. 1–28, 2022.

[17] W. Sun et al., "Online distraction detection for naturalistic driving dataset using kinematic motion models and a multiple model algorithm," in *Transp. Res. Part C, Emerging Technol.*," vol. 130, 2021, Art. no. 103317.

[18] M. Aljasim and R. Kashef, "E2DR: A deep learning ensemblebased driver distraction detection with recommendations model," *Sensors*, vol. 22, 2022, Art. no. 1858.

[19] N. Zangi, R. Srour-Zreik, D. Ridel, H. Chasidim, and A. Borowsky, "Driver distraction and its effects on partially automated driving performance: A driving simulator study among young-experienced drivers," *Accident Anal. Prevention*, vol. 166, 2022, Art. no. 106565.

[20] B. Li et al., "A new unsupervised deep learning algorithm for fine-grained detection of driver distraction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19272–19284, Oct. 2022.

[21] F.I. Rahmillah, A. Tariq, M. King, and O. Oviedo-Trespalacios, "Is distraction on the road associated with maladaptive mobile phone use? A systematic review," *Accident Anal. Prevention*, vol. 181, 2023, Art. no. 106900.

[22] 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.

[23] G. Gallego et al., "Event-based vision: A survey," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[24] M. S. Dilmaghani, W. Shariff, C. Ryan, J. Lemley, and P. Corcoran, "Control and evaluation of event cameras output sharpness via bias," *Proc. SPIE*, vol. 12701, 2023, pp. 455–462.

[25] R. Graca, B. McReynolds, and T. Delbruck, "Shining light on the DVS pixel: A tutorial and discussion about biasing and optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4044–4052.

[26] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "DSEC: A stereo event camera dataset for driving scenarios," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 4947–4954, Jul. 2021.

[27] P. Wzorek and T. Kryjak, "Traffic sign detection with event cameras and DCNN," in *Proc. Signal Process., Algorithms Architectures Arrangements Appl. Conf.*, 2022, pp. 86–91.

[28] C. Ryan et al., "Real-time face & eye tracking and blink detection using event cameras," *Neural Netw.*, vol. 141, pp. 87–97, 2021.

[29] 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.

[30] F. Becattini, F. Palai, and A. D. Bimbo, "Understanding human reactions looking at facial microexpressions with an event camera," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9112–9121, Dec. 2022.

[31] A.N. Angelopoulos, J. N. P. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein, "Event based, near eye gaze tracking beyond 10, 000 Hz," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 5, pp. 2577–2586, May 2021, doi: 10.1109/TVCG.2021.3067784.

[32] G. R. D. A. Moreira, "Neuromorphic event-based facial identity recognition," Master thesis, Universidade de Coimbra, Coimbra, Portugal, 2021.

[33] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[34] G. Lenz, S. H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Front. Neurosci.*, vol. 14, 2020, Art. no. 587.

[35] C. Yang, P. Liu, G. Chen, Z. Liu, Y. Wu, and A. Knoll, "Event-based driver distraction detection and action recognition," in *Proc. IEEE Int. Conf. Multisensor Fusion Integration Intell. Syst.*, 2020, pp. 1–7, doi: 10.1109/MFI55806.2022.9913871.

[36] L. Annamalai, A. Chakraborty, and C. S. Thakur, "EvAn: Neuromorphic event-based sparse anomaly detection," *Front. Neurosci.*, vol. 15, 2021, Art. no. 699003.

[37] Prophesee, "Data-preprocessing," Dec. 2022. Accessed: Mar. 08, 2023. [Online]. Available: https://docs.prophesee.ai/stable/metavision_sdk/modules/ml/data_processing/event_preprocessing.html

[38] J. Lemley, R. Corcoran, P. Kielty, M. Stec, and P. Toomey, "An introduction to the Xperi driving simulation environment: Hardware, software and data acquisition," in *Proc. Ir. Mach. Vis. Image Process. (IMVIP) Conf.*, 2023, doi 10.5281/zenodo.8212691.

[39] R. Benosman, S. H. Ieng, C. Clercq, C. Bartolozzi, and M. Srinivasan, "Asynchronous frameless event-based optical flow," *Neural Netw.*, vol. 27, pp. 32–37, 2012.

[40] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "HOTS: A hierarchy of event-based time-surfaces for pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1346–1359, Jul. 2017, doi: 10.1109/TPAMI.2016.2574707.

[41] L. Cordone, B. Miramond, and P. Thierion, "Object detection with spiking neural networks on automotive event data," in *Proc. Int. Joint Conf. Neural Netw.*, 2022, pp. 1–8.

[42] Prophesee, "Prophesee camera EVK3," Jan. 2022. Accessed: Mar. 08, 2023. [Online]. Available: https://www.prophesee.ai/event-based-evk-3/

[43] L. Berlincioni et al., "Neuromorphic event-based facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, 2023, pp. 4109–4119, doi: 10.1109/CVPRW59228.2023.00432.

[44] H. Bulzomi, M. Schweiker, A. Gruel, and J. Martinet, "End-to-end neuromorphic lip reading," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2023, pp. 4101–4108.
[45] U. Bissarinova, T. Rakhimzhanova, D. Kenzhebalin, and H. A. Varol, "Faces in event streams (FES): An annotated face dataset for event cameras," May 19, 2023, doi: 10.36227/techrxiv.22826654.v2.
[46] P. Kielty, M. S. Dilmaghani, C. Ryan, and J. Lemley, and P. Corcoran, "Neuromorphic sensing for yawn detection in driver drowsiness," in *Proc. 15th Int. Conf. Mach. Vis. (ICMV 2022)*, Jun. 2023, vol. 12701, pp. 287–294.
[47] P. Kielty, C. Ryan, M.S. Dilmaghani, W. Shariff, J. Lemley, and P. Corcoran, "Neuromorphic seatbelt state detection for in-cabin monitoring with event cameras," in *Proc. Ir. Mach. Vis. Image Process. Conf.*, Aug. 2023, doi: 10.5281/zenodo.8223905.
[48] 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.*, 2015, pp. 4489–4497.
[49] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5533–5541.
[50] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
[51] D. Liu, T. Yamasaki, Y. Wang, K. Mase, and J. Kato, "Toward extremely lightweight distracted driver recognition with distillation-based neural architecture search and knowledge transfer," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 764–777, Jan. 2023, doi: 10.1109/TITS.2022.3217342.
[52] L. Borawar and R. Kaur, "ResNet: Solving vanishing gradient in deep networks," in *Proc. Int. Conf. Recent Trends Comput.*, 2022, pp. 235–247.

WASEEM SHARIFF received the B.E. degree in computer science from the Nagarjuna College of Engineering and Technology, Bengaluru, India, in 2019, and the M.Sc. degree in computer science, specializing in artificial intelligence from the National University of Ireland Galway, Galway, Ireland, in 2020. He is currently working toward the Ph.D. degree with the University of Galway, Galway, under the IRC Employment Ph.D. Program. He is also a Research & Development Engineer with Xperi Inc. based in Galway. His research interests include machine learning for computer vision applications, with a particular emphasis on automotive driver monitoring applications

MEHDI SEFIDGAR DILMAGHANI received the B.Sc. degree in electronics engineering from the University of Tabriz, Tabriz, Iran, in 2012, and the M.Sc. degree in electronics engineering from the K. N. Toosi University of Technology, Tehran, Iran, in 2016. He is currently working toward the Ph.D. degree with the Department of Electrical and Electronics Engineering, University of Galway, Galway, Ireland, under the Hardiman scholarship. During the M.Sc. studies he had focus on electronic implementation of signal processing algorithms and wavelets. He is also a Research & Development Intern with Xperi Inc. His research interests include deep learning, computer vision, and event cameras.

PAUL KIELTY received the B.E. degree in electronic and computer engineering from the University of Galway, Galway, Ireland, in 2021. He is currently working toward the Ph.D. with the University of Galway and the ADAPT SFI research centre. His research focuses on deep learning methods with neuromorphic vision, with particular interest in driver monitoring tasks.

JOE LEMLEY received the B.S. degree in computer science and the master's degree in computational science from Central Washington University, Ellensburg, WA, USA, in 2006 and 2016, respectively, and the Ph.D. degree from the National University of Ireland, Galway, Ireland. He is currently a Principal R&D Engineer and Manager with Xperi Inc., Galway. His field of work is machine learning using deep neural networks for tasks related to computer vision. His research interests include computer vision and signal processing for the driver monitoring system.

MUHAMMAD ALI FAROOQ received the B.E. degree in electronic engineering from Iqra University, Karachi, Pakistan, in 2012, the M.S. degree in electrical control engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2017, and the Ph.D. degree from the National University of Ireland Galway (NUIG), Galway, Ireland. He is currently a Postdoctoral Researcher with NUIG. He is a Machine Learning Research Intern with Xperi Corporation. His Ph.D. research program was funded through the prestigious H2020 European Union (EU) Scholarship. His research interests include machine vision, computer vision, video analytics, machine learning, thermal imaging, and sensor fusion.

FAISAL KHAN received the B.S. degree in mathematics from the University of Malakand Chankdara, Pakistan, in 2015, the M.Phil. degree in mathematics from Hazara University, Mansehra, Pakistan, in 2017, and the Ph.D. degree from the National University of Ireland Galway, Galway, Ireland. His research interests include machine learning using deep neural networks for tasks related to computer vision, including depth estimation and 3-D reconstruction.

PETER CORCORAN (Fellow, IEEE) is currently the Personal Chair of electronic engineering with the School of Engineering, University of Galway (formerly known as National University of Ireland Galway), Galway, Ireland. He is also an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera redeye correction, and facial detection. He was a Co-Founder in several start-up companies, notably FotoNation, now the Imaging Division of Xperi Corporation. He has more than 600 technical publications and patents, more than 100 peer-reviewed journal articles, 120 international conference papers, and a coinventor of more than 300 granted U.S. patents. He is a member of the IEEE Consumer Electronics Society for more than 25 years. He is the Editor-in-Chief and the Founding Editor of *IEEE Consumer Electronics Magazine*.