IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 July 2023, accepted 16 July 2023, date of publication 20 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297500

RESEARCH ARTICLE

Real-Time Multi-Task Facial Analytics With Event Cameras

CIAN RYAN[1], AMR ELRASAD[1], WASEEM SHARIFF[1,2], JOE LEMLEY[1], PAUL KIELTY[2], PATRICK HURNEY[1], AND PETER CORCORAN[2], (Fellow, IEEE)
[1]Sensing Team, Xperi Inc., Galway, H91 V0TX Ireland
[2]Department of Electronic Engineering, College of Science and Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland

Corresponding author: Waseem Shariff (w.shariff1@nuigalway.ie)

This work was supported by Sensing Team, Xperi Galway. Additionally, authors received funding from the Irish Research Council under the grant number EBPPG/2022/17 with host institution, the University of Galway and Xperi Inc.

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

ABSTRACT Event cameras, unlike traditional frame-based cameras, excel in detecting and reporting changes in light intensity on a per-pixel basis. This unique technology offers numerous advantages, including high temporal resolution, low latency, wide dynamic range, and reduced power consumption. These characteristics make event cameras particularly well-suited for sensing applications such as monitoring drivers or human behavior. This paper presents a feasibility study on the using a multitask neural network with event cameras for real-time facial analytics. Our proposed network simultaneously estimates head pose, eye gaze, and facial occlusions. Notably, the network is trained on synthetic event camera data, and its performance is demonstrated and validated using real event data in real-time driving scenarios. To compensate for global head motion, we introduce a novel event integration method capable of handling both short and long-term temporal dependencies. The performance of our facial analytics method is quantitatively evaluated in both controlled lab environments and unconstrained driving scenarios. The results demonstrate that useful accuracy and computational speed is achieved by the proposed method to determining head pose and relative eye-gaze direction. This shows that neuromorphic facial analytics can be realized in real-time and are well-suited for edge/embedded computing deployments. While the improvement ratio in comparison to existing literature may not be as favorable due to the unique event-based vision approach employed, it is crucial to note that our research focuses specifically on event-based vision, which offers distinct advantages over traditional RGB vision. Overall, this study contributes to the emerging field of event-based vision systems and highlights the potential of multitask neural networks combined with event cameras for real-time sensing of human subjects. These techniques can be applied in practical applications such as driver monitoring systems, interactive human-computer systems and for human behavior analysis.

INDEX TERMS Event camera, neuromorphic sensing, driver monitoring system (DMS), head pose, eye gaze, facial occlusion.

I. INTRODUCTION

Driver monitoring systems (DMS) have become an essential part of the next-generation automobiles due to the primary cause of traffic accidents being driver error [1]. Driver monitoring systems can detect driver distraction, inattention, and fatigue, and can help prevent accidents from happening. In Europe, it is now mandatory for all new vehicles to be equipped with DMS [2]. Moreover, as we progress to higher levels of vehicle automation, understanding human behavior becomes increasingly important [3], [5]. There are several applications that can be beneficial for upcoming

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

advanced DMS systems, including eye gaze estimation and head pose estimation, which are important features for driver inattention, distraction, and fatigue detection [5].

Traditionally, cameras that capture visible or near-infrared (NIR) light are used for driver monitoring. However, these cameras operate at a fixed frame rate, have lower temporal resolution, capture all unwanted information in a scene, and have lower dynamic range or require additional lighting sources. Therefore, event cameras are being considered for advanced sensing paradigms in driver monitoring applications [6]. Event cameras generate a stream of asynchronous and independent events, which adapt their sampling rate to the scene dynamics, that is, motion. Event cameras are particularly suited to human or driver monitoring applications where high temporal resolution, low latency, and high dynamic range (140 dB) are required [6]. For example, eye blink [7], [8], fast pupil [9] and potential collision analysis.

This paper proposes a proof of concept for a deep learning approach to multi-task human facial analytics with event cameras. The proposed approach focuses on accomplishing the main tasks of driver monitoring, including eye gaze estimation and head pose estimation, simultaneously. Moreover, the network is trained entirely with simulated, synthetic event data [17], enabling us to train on non-event related, large-scale datasets. Gaze, pose, and occlusions estimation are considered core features of DMS [13], [16]. Traditional neural networks are trained and optimized for a single task, while driver monitoring systems consist of multiple and often related tasks. Thus, this paper proposes a multi-task convolutional neural network (CNN) to estimate head pose, eye gaze, and facial occlusions simultaneously. In a multi-task learning framework, a neural network is trained to predict multiple related tasks and shares low-level feature maps between complementary tasks. This has the advantages of improved efficiency, performance, and natural regularization of shared layers [15], [16]. Facial occlusions, such as masks or glasses, may obstruct important facial features, or may require special processing or attention. Therefore, it is often desirable to identify such occlusions for downstream applications that require knowledge of or can be hindered by such obstructions. The proposed deep learning approach addresses these issues by using event cameras and integrating them with AI-based image pipelines, to further benefit advanced driver assistance systems (ADAS) [10]. Moreover, event cameras are privacy-preserving, which makes them attractive to industry and consumers alike [11], [57].

Although event cameras offer several advantages, they are not suited for the detection or analysis of stationary objects [6]. In a driving setting, often the driver is a stationary object in a dynamic scene. Thus, this paper presents an event integration method to handle short-term and long-term periods of no motion. Specifically, a leaky time surface is employed to overcome short-term effects [18]. In addition, a region of interest (ROI) event threshold for long-term motionlessness is presented. This threshold determines the waiting time between face detection and facial analysis inference and is a function of the number of incoming events within the face region. It acts as an inference trigger enabling the multi-task network to respond only to face motion.

The applicability of event cameras in human or driver monitoring has not been extensively investigated in previous research [8]. This paper presents a significant contribution by providing a comprehensive proof-of-concept implementation of a multi-task facial analysis network specifically designed for event cameras. Note that certain components of this work were originally documented in [29], and [30]. This paper provides additional detail together with a comprehensive evaluation of our framework including test and validation on a dataset acquired with a recent model of event camera. Further the paper demonstrates the feasibility of event-based vision systems and highlight the potential of multitask neural networks combined with event cameras for real-time sensing of human subjects.

Key contributions of this paper are:

1) An innovative event-based multi-task facial analytics framework designed to efficiently handle end-to-end training and testing processes.
2) A set of experimental validations of this framework with a focus on the tracking of head pose and eye-gaze demonstrating its potential for real-time sensing of human subjects.
3) The use of synthesized event-stream data from conventional video datasets for training the network, and the validation of the use of the data by testing it on real event camera data.
4) Introduction of a novel algorithmic technique for event stream data, combining leaky integration and ROI motion thresholding to mitigate interference from global subject motions.

The remainder of this paper is broken down into the following sections. Section II presents a review of the literature related to event-based driver monitoring systems and conventional multi-task learning. Section III describes the dataset used in the study, including its distribution. Section IV outlines the methodology used in this research, which includes a novel event representation, network architecture, and training details. Section V covers the experiments and results obtained in this study, including results based on training with the synthetic event dataset, real event data, and performance comparison with other modalities and driver monitoring applications. Section VI has a brief discussion of the entire findings of the paper. Finally, the paper concludes with a summary of the findings and their implications in Section VII.

II. RELATED WORK

The event-based driver monitoring system and the multi-task learning and facial analysis are the two developments that this study primarily focuses on. The associated literature is presented below.

A. EVENT-BASED DRIVER MONITORING SYSTEM

Event cameras offer several significant advantages over conventional cameras, including high temporal resolution, high dynamic range, low power consumption [6], [49]. They have been applied to applications such as robotics and surveillance. Chen et al. [20] provide a comprehensive overview of event cameras and their applicability to autonomous driving tasks. Moreover, several other studies have also explored the effectiveness of event cameras for autonomous driving [19], [21], [22], [23]. To date, limited attention has been given to the applicability of event cameras for human monitoring and facial analysis.

Ryan et al. [8] present a face and eye detection recurrent CNN and blink detection algorithm for driver monitoring purposes. Several other studies employ non-deep-learning methods. Lenz et al. [7] also present a face detector and tracking using the temporal signature of an eye blink. Liu et al. [24] present another approach to efficiently detect faces with event cameras. The authors use event 3D tensors with a lightweight translation-invariant backbone to extract multi-scale features. Authors in [25] proposed a face identity recognition with event cameras. Authors in [26], present real-time gaze tracking with eye segmentation. To more accurately measure the eye gaze, authors emulate events using near-eye camera frames. Chen et al. [27] detect eye and mouth motions from events and extract drowsiness-related features for driver monitoring. Moreover, [31] also describe neurobiometric method, a bio-metric authentication system using event based eye blinks. Anastasios et al. [9] present a hybrid event and frame-based method of tracking eyes. Authors in [28], suggest an event-based dataset and benchmark study on three different components of distracted driving: driver fatigue, visual attention, and hand movements. There have been few studies which explore the micro-movements of the driver including the expression [11] and driver distraction [24]. Moreover, the latest developments in event-based vision have the potential to enhance advanced driver assistance systems and driver monitoring systems. These advancements encompass various areas such as person detection [50], human pose estimation [51], event-based facial expression recognition [52]. This research draws inspiration from various aspects explored in previous studies [29], [30]. It delves into several areas, such as neuromorphic algorithms, object detection using event cameras, and techniques for generating textural images based on event count decay factor and net polarity.

B. FACIAL ANALYSIS AND MULTI-TASK LEARNING

Previous research has explored the integration of real-time driver monitoring systems. For instance, in [54], the authors introduced a real-time driver drowsiness detection approach. Additionally, several other studies have investigated methods that could prove beneficial for Driver Monitoring Systems (DMS), YOLO-face [55], and an interaction system designed to measure human attention levels [56].

Multi-task learning is used to train multiple related tasks in a unified framework with the aim of improving learning efficiency, reducing memory and inference time and improving performance relative to individual task-specific networks [32], [33]. Several studies have adopted multi-task learning for facial analysis in recent years. Hu et al. [34] address the problem of subtle expression recognition by training a multi-task network to estimate emotion and facial landmarks. The rationale is that subtle expression changes often coincide with facial landmark movements. Ranjan et al. [35] propose HyperFace, a multi-task CNN trained for five facial analysis tasks. The authors present two novel CNNs with truncated AlexNet and ResNet as backbones. Recently, Yang et al. [16] proposed a multi-task (MT) network for the pose, gaze and landmark estimation in a driver monitoring setting. While multi-task learning is widely employed in the literature it has not yet been applied to asynchronous imaging techniques such as event camera systems. One of our core contributions is to adapt the multi-task approach to enable a multi-task training framework to be realized with event camera vision systems.

III. DATASETS

A. TRAINING DATA

A major impediment to event-based machine learning research is the lack of labelled public datasets. However, recent developments in synthetic event simulators have unlocked the use of datasets collected from alternative, frame-based cameras. This paper utilizes V2E, an open-source event simulator [17] to generate synthetic event streams from fixed labelled frames. The simulator calculates log-intensity changes beyond a predefined contrast threshold between successive frames to generate events [17]. V2E also models realistic temporal and leak noise. A large near-infrared (NIR) dataset was collected with an OptiTrack motion sensor to annotate head pose, eye gaze and facial occlusions. The data was acquired in a laboratory environment on a driving simulator. These NIR sequences are converted to event streams using the above-mentioned event simulator. The locally acquired NIR dataset consists of approximately 21,000 training images, 4,600 validation images and 2,800 testing images with a resolution of 224×224. Subject ages span from 18 to 55. Occlusions comprise eye (i.e., glasses) and mouth (i.e., mask) occlusions. Overall, approximately 50% of the dataset contains eye occlusions and 30% contains mouth occlusions. Figure 1 shows the distribution of head pose and eye gaze for yaw and pitch angles in our training dataset. The test dataset exhibits a similar distribution.

Each image is first subjected to a series of random augmentation and transformations that simulate a video sequence with homographic camera motion with six degrees of freedom (6DOF). We simultaneously change the annotation using the same transformation matrix for each image. As a consequence, a video sequence with frame-by-frame annotations is produced. Further as mentioned above we use V2E to

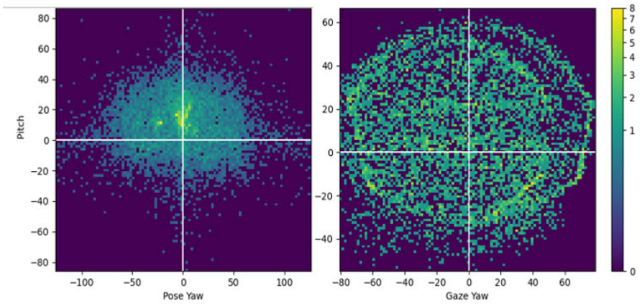


FIGURE 1. Distribution of head pose and eye gaze labels along x and y axes.

simulate events from these generated sequences. Inspired by [8] we adopt a similar approach to train our proposed CNN network. With the help of this effective approach, we train CNNs to work in event space without the use of intermediary intensity representations. This approach enables the network to effectively generalize to real-time cameras.

The contrast threshold (CT) is a central parameter for event cameras. It determines whether the change in light intensity is large enough to generate an event. This parameter was tuned in event simulators to simulate NIR frames. Contrast thresholds are typically samples from $N(0.18, 0.03)$ [37]. Others have found improved generalization capabilities with wider-ranging thresholds between 0.2 and 1.5 [38], [39]. However, these results are all based on RGB images, and this paper simulates events from near-infrared data. This yields unique challenges. We found that intensity changes in the infrared spectrum require a significantly lower contrast threshold due to the naturally lower brightness levels. As a result, we sample CT from $N(0.12, 0.03)$.

B. TESTING DATA

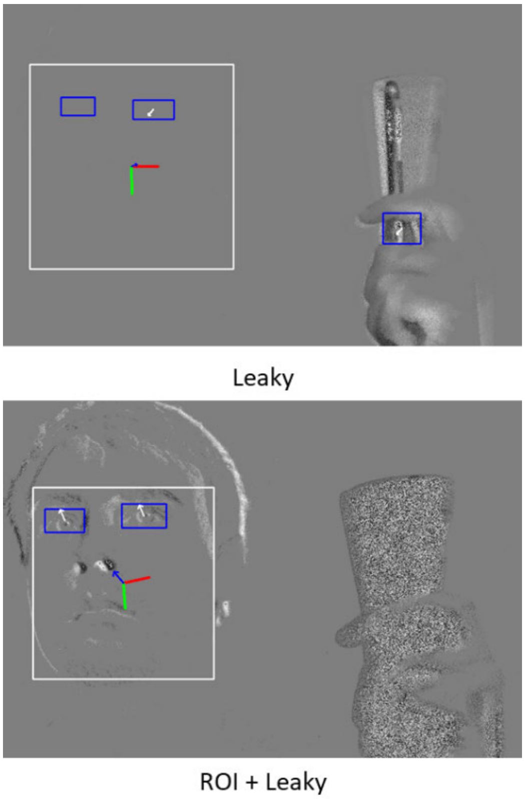
The performance of the multi-task driver monitoring network is tested quantitatively on both synthetic and real events. First, performance is reported on both validation and test event-simulated datasets. Second, the performance of the head pose is tested with event data captured from a high-resolution Prophesee event camera with a resolution of 1280 × 720 pixels. The data acquisition setup consisted of 5 subjects who were instructed to drive in a driving simulator. The data was collected with informed consent. Samples from each acquisition were taken for testing purposes. Head pose annotations were collected continuously using an OptiTrack motion sensor system. Lastly, qualitative results are demonstrated in an unconstrained driver monitoring environment (see supplementary material for video).

IV. METHODOLOGY

This section details the proposed methods used to construct the human/driver monitoring and facial analysis system and is split into 4 subsections: A) The formulation of the event-based representation supporting short-and long-term periods of no motion or inactivity, B) The design of the multi-task CNN architecture for the head pose, eye gaze and occlusion detection, C) Facial analytic inference with face and eye detection networks: a two-network approach and D) Details surrounding network training and inference.

A. EVENT REPRESENTATION

Event cameras capture light intensity changes caused mostly by moving objects. As a result, they are particularly suited to the detection, tracking and analysis of moving objects. However, they are not suited to recognizing slow-moving or static objects and the appearance of an object is highly dependent on the relative motion [6]. To leverage event cameras as standalone human monitoring systems, methods to handle stationary or static human operators are needed. Limited attention has been given to this problem. Zanardi et al. [40] point out that an object is only detectable in RGB and event domains with motion. Maqueda et al. [21] experienced this problem and removed data collected from a vehicle with speeds under 20km/h.


Leaky

ROI + Leaky

FIGURE 2. Comparison of leaky time surface with and without the ROI inference trigger.

Inspired by [29], and [30], two factors contribute to this phenomenon: 1) the absence of motion in the ROI and 2) the faster motion of other irrelevant objects. This paper proposes a combination of leaky time surface and ROI-based inference triggering to handle short and long-term periods of stationarity. Figure 2 below demonstrates the results of the proposed method.

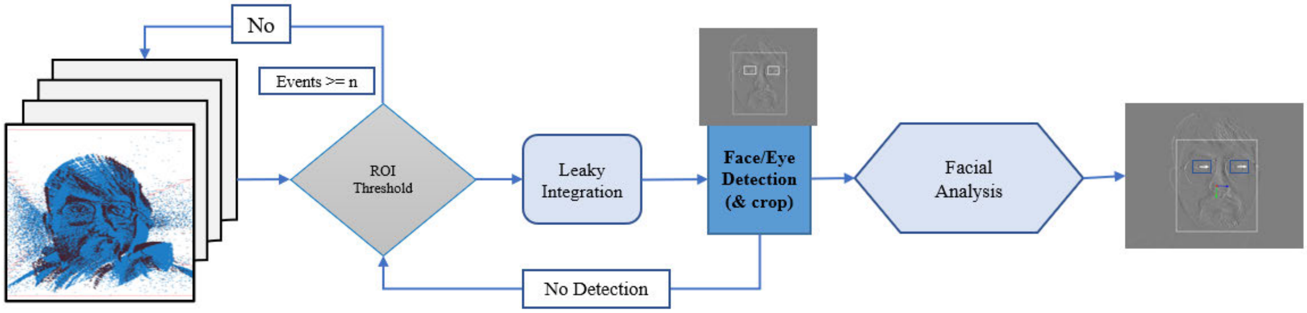


FIGURE 3. Process pipeline for mapping events to image space and processing with the proposed neural networks.

In our method, we utilize time surfaces to process events in the neural network, as proposed in [6]. Time surfaces are used to retain short-term memory of recent events and are constructed by recording the timestamps of the most recent event in each pixel, with the polarities of each event considered separately. Larger intensity values in the time surface reflect more recent motion at that pixel location [6]. An exponential leaky time surface (ETS) is used, which emphasizes recent events over past events [6]. The leaky time surface is updated incrementally with the associated polarity of the recorded event at the corresponding pixel, and it decays over time based on a predefined factor that corresponds to the time elapsed between the most recent and previous events. In this way, the time surface holds a memory of recent events that are decayed as a function of time [18], [41]:

$$ETS_i(x, y, t) = P_i(x, y).e^{\frac{T_i(x,y)-t}{\tau}} \tag{1}$$

where x, y, t represent the x, y coordinates and timestamp of event i. T_i is the time surface that maps the time of the previous and most recent events to the pixel location P_i. As a result, the pixel value is decayed more if the time difference between the current and previous event is larger i.e., as $T_i(x, y) - t$ approaches -τ. τ is set to 100ms in this paper. According to figure 3, if the number of event samples greater than n, then the associated events are taken into account. The n is assumed for the sake of this study to be 20k inside the face region.

The leaky time surface acts as a form of feature map, encoding temporal information from events into intensity values that represent the recency of motion at different pixel locations. It is then used as input to the face detection network, serving as a representation of the motion information in the ROI for face detection. The ROI is determined based on the inference triggering mechanism that combines the leaky time surface and ROI-based threshold. The face detection network performs face detection based on the encoded temporal information from the time surface. An ROI inference trigger is presented to overcome some limitations of the leaky time surface, which preserves information over shorter time periods limited by the parameter τ. To account for longer periods of absent object motion, an ROI-based threshold is proposed as a method to trigger the next instance of detection and face analysis based on apparent changes to the human operators' state. By combining leaky time surfaces and ROI-based inference triggering, the proposed method effectively handles the disproportionate motion between the object of interest (e.g., the face) and the background. It captures and represents the motion information specific to the ROI, enabling accurate detection and analysis even in scenarios where the background exhibits faster or different motion patterns. Faster motion generates more events, and thus the number of events is proportional to velocity. Therefore, this paper adopts an inference trigger based on the number of incoming events within the ROI, reflecting apparent changes to the state of the human operator due to movement. The trigger ensures a minimum number of face events before further facial analysis. Figure 2 illustrates the comparison between leaky time surface (LTS) with and without the ROI inference trigger. The figure showcases the impact of the ROI trigger on the recognition of facial features using n event samples. The ROI trigger, when applied to the face areas, enhances the LTS model's ability to accurately identify and analyze facial features.

The method establishes a waiting time between the next instance of detection and facial analysis. Similar to [42], a threshold is set as a number of incoming events proportional to the size of the ROI set to $\mu \times height_{roi} \times width_{roi}$. μ is set to 0.10 in our experiments. This both increases efficiency (only perform inference when the driver moves) and maintains performance (network input is not entirely sparse). In [42], and [43], a long-term object tracking framework is proposed. A key parameter defined in their method is the "waiting time" between two instances of classification. Inspired by [42], this paper adopts an ROI-based event threshold to determine the waiting time between the next instance of detection and facial analysis. In [42], and [43], manual identification of the ROI is required. A local sliding window method is then applied to the small ROI. Classification is performed on each window given enough incoming events. Unlike [42], the method in this paper determines the waiting time for both face detection and facial analysis together. This improves detection at the next phase and the performance of facial analytics. Moreover, this method does not require initial manual initialization. This paper extends the method originally proposed in [42], and [43] to determine human movements and trigger the next instance of facial analysis and object detection.

The processing pipeline can be seen in Figure 3. In this figure it can be observed, the block contains event streams generated from NIR images using the V2E method. These event streams are then fed into a series of processing steps that aim to extract useful information from the data. Specifically, we first pass the event streams through an ROI threshold trigger, which identifies ROI in the images. The resulting output is then fed into a Leaky integrator, which smooths the event stream to reduce noise and enhance the signal. The ROI-based events are utilized as inputs to the leaky integrator. This component processes the incoming events over time and generates the leaky time-surface representation. The leaky integrator incorporates the new events into the existing representation while exponentially decaying the impact of past events. Next, a face/eye detector is applied to identify the face and eye regions within the images. Finally, we propose a novel multi-task network that utilizes the identified facial regions to perform head pose estimation, eye gaze estimation, and facial occlusion detection. This comprehensive approach allows us to extract a wealth of information from the event streams, providing insights into the subject's facial expressions, eye movements, and head orientation. The use of event-based data and the proposed multi-task network have significant implications for facial analytics. By leveraging event streams generated from NIR images, we can capture fine-grained changes in the image data, resulting in highly accurate and efficient facial analysis. The multi-task network provides a robust and versatile framework for performing multiple tasks simultaneously, enabling us to extract more information from the data than traditional methods. Overall, this project represents a significant step forward in the field of facial analytics and has the potential to revolutionize how we understand and analyze human behavior.

B. NETWORK ARCHITECTURE

The proposed system, inspired by [8] and [15], is a two-stage framework consisting of two CNN. The first CNN locates and tracks the face and eyes, while the second CNN estimates head pose, eye gaze, and occlusions in a multi-task learning framework.

1) FACE AND EYE DETECTION

In this research, the authors adopted a novel face and eye detection network developed by Ryan et al. [8]. This network is used to identify and track faces and eyes, which are then used as ROI for further analysis. The network architecture proposed in [8] is specifically designed for driver monitoring and has been trained using the Helen RGB-based dataset. This dataset was collected in the wild, unlike the authors' dataset, which was collected in a lab setting. To simulate near-infrared (NIR) images, the authors used V2E. The [8] network is a recurrent convolutional neural network (RCNN) that utilizes a modified version of YOLOv3-tiny, a state-of-the-art object detection algorithm. It also has a fully convolutional gated recurrent unit that maintains temporal memory about the location of the face and eyes. This feature allows for accurate face detection even when there is limited face information available. During training, the authors used an augmentation technique to ensure that the network can maintain information during periods of prolonged driver stillness. During inference, the input size to the network is 512×288, based on a down-sampled input from the same Prophesee (1280×720) camera used in this study. The use of this network is similar to how the face detector library works, where it delivers face bounding boxes that are used to crop the face regions. These face crops are then used as inputs to the multi-task network for facial analysis in the next stage. To gain further insights into the architecture and training of the network mentioned in [8], readers are encouraged to refer the original paper. Our work is based on this research, and the original paper provides comprehensive details on the subject.

2) FACIAL ANALYSIS

The facial analysis network proposed in this paper is a multi-task CNN that is designed to detect head pose angles, eye gaze angles, and face occlusions. This is a significant improvement over using a single network for each task, as multi-task networks can share feature representations in lower layers, leading to improved efficiency, computational cost, and performance. Its capabilities render it an influential instrument for real-time applications, particularly when integrated into the onboard sensor suite of advanced driver assistance systems.

The proposed network architecture, shown in Figure 4, consists of a series of shared parameters and individual task-specific layers. The input to the network is a face crop with a resolution of 224×224. The outputs of the fully connected layers differ per task, with only two columns for gaze (yaw, pitch) and occlusions (eyes, mouth). Each block in Figure 4 displays the type of layer, kernel size, number of out-feature maps, and stride. A Leaky-Relu activation function is used at each convolutional layer. A stride of 1 is assumed if not explicitly shown. The first 5 layers are shared among all tasks, shown in grey, with a skip connection at layers 3 to 5, and other skip connections represent concatenation. The proposed network is based on the All-in-One facial analysis network proposed in [15], but it differs in several key areas. Smaller kernel sizes are used in the early layers to capture smaller, more complex features such as pupil information. Task-specific convolutional layers are also added, and less fully connected layers are used for each task. The authors of the paper split tasks into subject-dependent and subject-independent. Layers 1, 3, and 5 are utilized for subject-independent tasks [19], [35]. Similarly, these layers are combined for pose and occlusion detection, as these tasks are strongly related. The existence of occlusions can be used to better estimate pose, and gaze estimation requires more detailed features of pupils. The extraction of pose features in

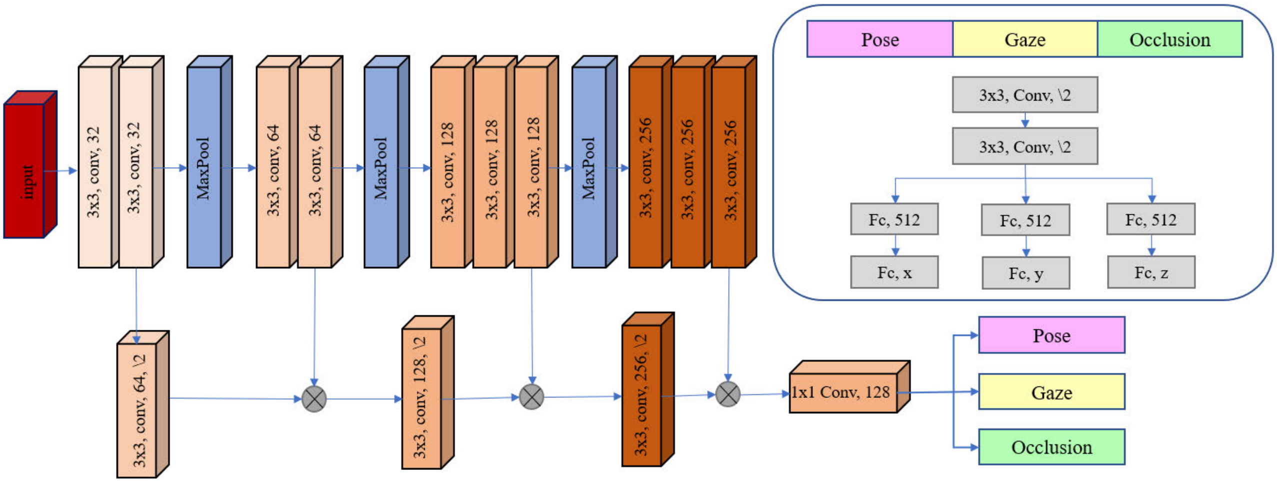


FIGURE 4. Multi-Task Network Architecture. Left: Overall Architecture. Top-Right: Structure for each specific task.

the shared layers supports improved eye gaze estimates given the relationship between these tasks.

Multi-task learning is a popular area of research in machine learning, where multiple related tasks are learned simultaneously by sharing parameters and representations across them. It has been shown to improve the performance of individual tasks and reduce training time and computational cost. In the context of facial analysis, multi-task networks have been used to detect facial expressions, age, gender, and other attributes, in addition to the tasks addressed in this paper. The proposed network has several potential applications, especially in the field of advanced driver assistance systems. For example, it can be used to detect driver distraction, drowsiness, or impairment by analyzing head pose and gaze angles. It can also be used for face recognition, which can enhance the security of the vehicle and prevent unauthorized access. Moreover, it can be applied to the field of human-computer interaction, such as in gaming, virtual reality, or teleconferencing, where accurate gaze estimation is critical for user experience. Moreover, the facial analysis network proposed in this paper is a multi-task CNN that is designed to detect head pose angles, eye gaze angles, and face occlusions simultaneously. It has several advantages over using a single network for each task, including improved efficiency, computational cost, and performance. The proposed network has several potential applications in the field of advanced driver assistance systems, human-computer interaction, and other areas where facial analysis is critical.

Algorithm 1 Multi-Task Facial Analytics Inference

```
 Initialize Leaky_int ← LeakyIntegrator()
 Initialize time_surface ← $P_i(x, y).e^{\frac{T_i(x,y)-t}{\tau}}$
 Initialize event_window_iterator ←
 FIXEDSIZEEVENTREADER(coords) or
FIXEDDURATIONEVENTREADER(duration) or
FIXEDSIZEROIREADER(coords) eye_xc, eye_yc ← []
missing_face ← 0
for events in event_window_iterator do
    time_surface_grid ← TIME_SURFACE(events)
    input ← LEAKY_INT(time_surface_grid)
    # Model Detect for Face & Eye Coordinates
    output_det ← MODEL_DETECT(input)
    face_coords, eye ← non_max_suppression
(output_det)
    if face_coords is None then
        return missing_face += 1
        if missing_face > 0 then
            break
        end if
    else
        continue
    end if
    # Multi-task Facial Analytics (AIO)
    NEW_INPUT ← extract_events_within(face_coords,
eye)
    OUTPUT_AIO ← MODEL_AIO(NEW_INPUT)
    yaw, pitch, roll, eye_xc, eye_yc, occl ← OUTPUT_AIO
end for
```

C. TWO-NETWORK APPROACH FOR FACIAL ANALYTIC INFERENCE

Algorithm 1 presents a pseudo code representation of the methodology used for multi-task facial analytic inference. This pseudo code outlines the step-by-step process employed in the inference phase, offering a clear and concise depiction of the methodology's implementation. The provided pseudo code outlines the methodology for multi-task facial analytic inference using two networks. The first network focuses on face and eye detection. It starts by initializing a LeakyIntegrator and defines an event window iterator based on fixed-size or fixed-duration or ROI based events. The code then iterates over the event windows, creating a time surface grid and passing it through the LeakyIntegrator. The resulting input is fed into the face and eye detection model, and non-maximum suppression is applied to obtain the coordinates of detected faces and eyes. The second network, referred to as AIO (All-In-One), handles tasks such as pose estimation, gaze tracking, and occlusion detection. The input to this network consists of events within the regions defined by the face and eye coordinates obtained from the previous step. The AIO model processes this input, producing the corresponding outputs. Throughout the process, the code tracks and stores eye center points for visualization, and also smooths the outputs of the AIO network for better visualization. It includes visualizations of face crops and displays the outputs, such as head pose, eye gaze direction, and occlusion classification. In summary, this pseudo code represents the implementation of a multi-task facial analytic inference system using two networks. The first network focuses on face and eye detection, while the second network performs pose estimation, gaze tracking, and occlusion detection. The code includes various steps for data preprocessing, network inference, output tracking, and visualization to provide comprehensive facial analytic results.

D. TRAINING DETAILS

The multi-task network was trained with pre-made events from NIR frames (approximately 21,000 images and validated with 4,600 images and training with 2,800 images). The network was implemented with PyTorch. An ADAM optimizer was used with a learning rate of 3×10^{-3} reduced by a factor of 0.8 every 20 epochs for 120 epochs. Weight decay of 3×10^{-2} was applied. Our multi-task training requires 3 individual loss functions for each task. For eye gaze and head pose, mean squared error (MSE) loss was used. Cross entropy loss was used for occlusion classification. The network was implemented on a Nvidia GeForce RTX 2070 GPU. The average processing time for accumulated sub-sample of events is 2.4 milliseconds for the multi-task network. Thus, the multi-task network can operate at an equivalent event framerate of over 400 fps. The actual required framerate naturally fluctuates based on scene dynamics as per the output of the event camera. The network contains 4.9 million parameters. The average time for the face and eye detection network is 9 milliseconds, comprises of 12.8 million parameters and contains approximately 5.82 GFLOPS.

V. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed multi-task network and driver monitoring system, a series of tests were conducted using various types of data. Firstly, synthetic event datasets were used to assess the network's ability to accurately detect and classify different events, such as head pose angles, eye gaze angles, and face occlusions. This allowed the authors to validate the accuracy and robustness of the system under controlled conditions. Secondly, real event data from a high-resolution event camera was used to further evaluate the network's performance. This involved collecting data from various driving scenarios, such as different lighting conditions and driving speeds, to assess how well the network performs in real-world scenarios. Lastly, a qualitative evaluation was conducted in a car setting to test the driver monitoring system's effectiveness in real-world scenarios. This involved installing the system in a car and collecting data while a human driver was driving the car. The collected data was then evaluated to assess the accuracy and effectiveness of the driver monitoring system.

Overall, these tests allowed the authors to thoroughly evaluate the proposed multi-task network and driver monitoring system's performance under a range of conditions. By using synthetic and real event data, as well as conducting a qualitative evaluation in a car setting, the researchers were able to validate the network's accuracy and robustness in a variety of scenarios. This approach helps to ensure that the system is reliable and effective in later detecting driver inattention, fatigue, or distraction, using the head-pose, eye-gaze and facial occlusions which are essential for advanced driver assistance systems to ensure safety on the roads.

Evaluation Metrics: In this study, we evaluate the accuracy of estimating head pose and eye gaze by calculating average absolute errors in degrees, whereas occlusion is reported in terms of detection accuracy. Recent advancements in CNN [44], [45], [46], [47] have allowed for direct regression of head pose using Euler angles. However, training neural networks to predict angles solely through a single regression loss function can be difficult. To overcome this challenge, we adopt a technique called Dual-loss Block (DLB) [16], which breaks down the pose estimation task into two parts: pose classification and pose coarse-to-fine regression. DLB combines two types of loss functions: classification loss and regression loss. Classification loss helps the network quickly converge to a small boundary, while regression loss helps the network learn the residual information to make more accurate predictions. We use cross entropy loss for classification and mean squared error for regression. The head pose and eye gaze outputs are calculated independently, but both utilize the same approach. To handle occlusions, we employ a straightforward cross-entropy loss for classifying them.

TABLE 1. Performance results on synthetic NIR-to-Event validation and training datasets. Pose and Gaze is in degrees. Occlusions are classification accuracy.

Synthetic Event Data Testing			
		Valid	Test
Pose	Yaw	2.84	3.60
	Pitch	3.67	4.49
	Roll	2.22	2.72
Gaze	Yaw	7.93	7.92
	Pitch	6.45	6.07
Occlusion	Eye	99.1%	99.0%
	Mouth	99.5%	99.6%

A. RESULTS ON SYNTHETIC DATA

The performance of the multi-task network on synthetic datasets has been evaluated and the results are presented in Table 1. The results indicate that the head pose estimation

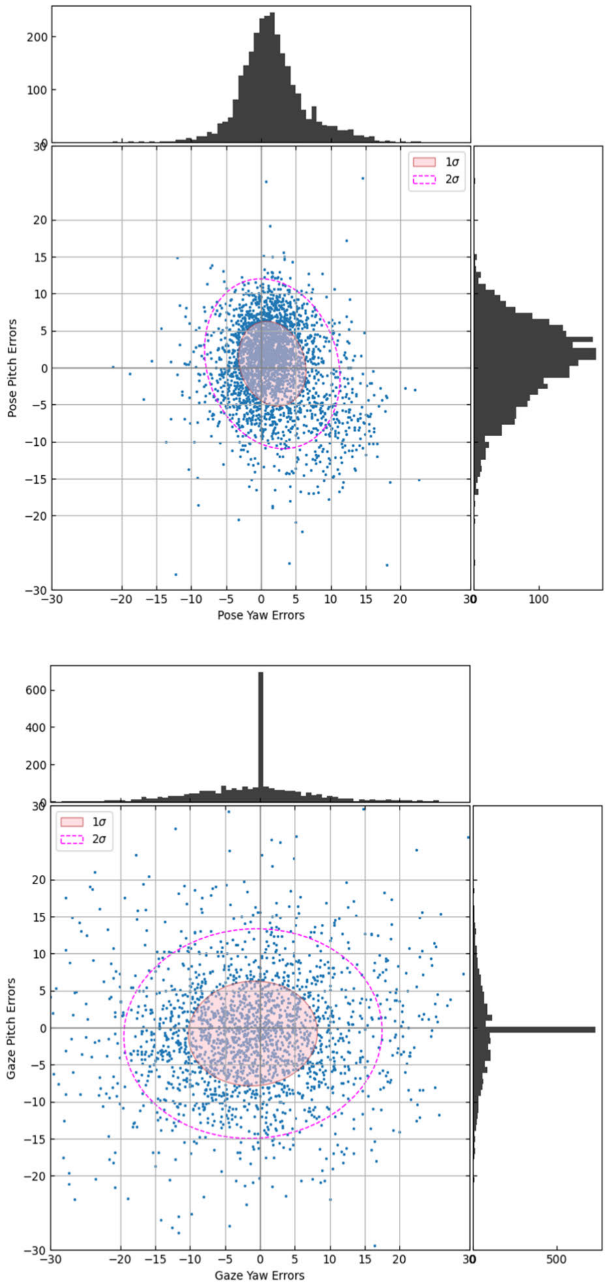


FIGURE 5. Distribution of error for head pose (top) and eye gaze (bottom) on the test dataset.

error is close to or below 5 degrees for both validation and testing datasets, which demonstrates the effectiveness of the multi-task approach. The average eye gaze prediction errors are also below 10 degrees for both datasets, despite the fact that eye gaze estimation is a more complex task relative to head pose, resulting in larger errors. To gain a better understanding of the distribution of errors, a 2D scatter plot is presented in Figure 5, which displays the distribution of errors for yaw and pitch for both head pose and eye gaze. It is worth noting that these results are based solely on the testing dataset. The scatter plot is accompanied by ellipses that represent 1 and 2 standard deviations from the mean. The outer ellipse contains approximately 95% of the data, whereas the inner ellipse contains 68%. Additionally, 1D histograms are displayed on the sides of each error scatter plot to aid in visualizing individual distributions. The results obtained from the synthetic data indicate that the multi-task approach can effectively estimate head pose and eye gaze with event cameras, even in the presence of occlusions.

TABLE 2. Mean absolute error in degrees for head pose on BIWI simulated events.

	BIWI [48]				
	Modality	Yaw	Pitch	Roll	MAP
FSA-Net [44]	RGB	4.50	6.08	4.64	5.07
HPE [45]	RGB	4.80	6.18	4.87	5.28
QuatNet [46]	RGB	3.97	5.62	3.92	4.50
FDN [47]	RGB	3.78	5.61	3.88	4.42
2DHeadPose [53]	RGB	3.81	3.78	2.73	3.44
AIO-Net(our)	Event	8.02	9.86	8.6	8.80

B. PERFORMANCE COMPARISON WITH OTHER MODALITY

It's worth noting that there has been a lack of recent research on event-based head pose estimation, which highlights the importance of our proposed network. In this study, we evaluated the performance of our trained model by comparing it with other existing models using the BIWI [48] dataset. We simulated events from visible images using 50 successive sequences with the 24 subjects available in the dataset [48]. The results are presented in Table 2, the results of the proposed network in event-based head pose estimation were almost 2 times worse than the state-of-the-art due to various reasons. Firstly, the face crops in the evaluation dataset had a lower resolution than the 224 × 224 input, which led to the need for up-sampling, resulting in a loss of quality. Secondly, the contrast threshold used could have been adjusted to improve the accuracy. Finally, there is potential for improvement by experimenting with different simulation settings and incorporating more events in the input. Despite these limitations, the proposed network still shows promising results in event-based head pose estimation and provides a foundation for future research in this field. The study emphasizes the need for further investigation to improve the accuracy of event-based facial analysis techniques.

C. RESULTS ON REAL EVENT DATA

The multi-task network is put to the test on real event data collected from a high-resolution Prophesee event camera with a resolution of 1280 × 720. The results of the head pose estimation for five different test subjects are presented in Table 3 below. However, it should be noted that eye-gaze and facial occlusion labelling were not available for these recordings due to the constraints of the Ground-Truth-motion sensor technology that was used. The results of the head pose

TABLE 3. Mean absolute error in degrees for head pose on real event camera data.

Real Event Data Testing				
	Yaw	Pitch	Roll	Mean
Baseline (Pretrained)	3.60	4.49	2.72	3.60
Sub-1	3.28	6.03	5.37	4.89
Sub-2	4.91	7.1	4.28	5.43
Sub-3	5.03	6.98	3.86	5.29
Sub-4	5.61	6.3	2.96	4.95
Sub-5	2.69	5.02	2.36	3.35
Average	4.30	6.29	3.77	4.78

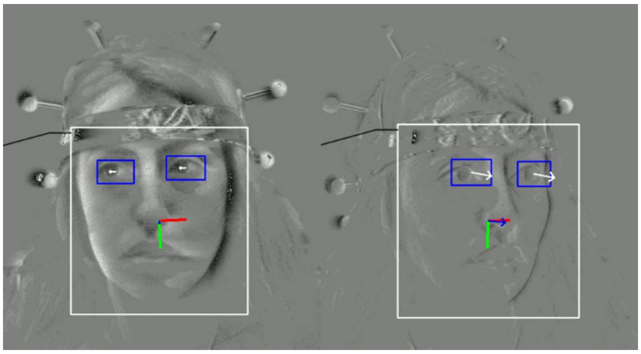


FIGURE 6. Qualitative results of testing on real event data.

estimation are reported as a mean absolute error in degrees, and they are found to be comparable to those obtained from the synthetic event data in the previous section. This indicates that the multi-task network has generalization capabilities and can effectively estimate head pose in real event data, even though it was entirely trained using synthetic events. To further illustrate these results, qualitative examples are provided in Figure 6. The figure displays red, green, and blue arrows, which correspond to pose vectors, and white arrows in the eye regions that correspond to gaze vectors. These results provide valuable insights into the performance of the multi-task network on real event data, which is crucial for the practical implementation of this technology in real-world scenarios. The results of the head pose estimation on real event data demonstrate that the multi-task network can effectively estimate head pose, even in the absence of eye-gaze and facial occlusion labelling, and has the potential for various applications such as robotics and human-computer interaction where accurate head pose estimation is crucial.

D. DRIVER MONITORING APPLICATIONS

Driver monitoring is a safety-critical human-operator monitoring system that requires accurate facial analysis at all times. This paper proposes a human monitoring system designed specifically for driver monitoring, leveraging the advantages of event cameras such as temporal resolution, response to motion, and high dynamic range, while addressing the limitations of human stillness and stationarity. To demonstrate the effectiveness of this system, it is applied to an unconstrained driver monitoring scenario. To provide qualitative analysis, a single subject is monitored during a single trip. Figure 6 displays a screenshot of the trip, showcasing the results obtained from the driver monitoring system. For further analysis, a video recording of the trip can be found in the supplementary material. Algorithm 1 shows the inference run of proposed multitask neural network. The proposed system employs a multi-task network, leaky integration event representation, and ROI-based inference trigger to enhance the accuracy of facial analysis. These techniques exploit the unique features of event cameras, making it well-suited for driver monitoring applications. The driver monitoring system presented in this paper has a wide range of practical applications, such as improving safety in the automotive industry. By continuously monitoring drivers and detecting any signs of fatigue or distraction, this system can alert drivers to take necessary breaks, thereby reducing the risk of accidents. Overall, the results obtained from the driver monitoring system indicate its potential to significantly improve the safety of road transport.

VI. LIMITATIONS

The performance of the trained model on the BIWI simulated dataset was not state-of-the-art, indicating the need for further optimization or exploration of alternative approaches. Additionally, it is important to note that the proposed approach and evaluation primarily concentrated on driver monitoring scenarios. Generalizing the findings to other facial analytics domains or broader applications would benefit from additional investigation. Furthermore, although the proposed approach demonstrated impressive results, further fine-tuning or optimization may be necessary to enhance its performance on specific datasets or in diverse real-world scenarios. One specific area that could be improved is the face detection algorithm used in this research. While the adopted face detection network showed promising results, there is still room for improvement. Exploring alternative algorithms or incorporating state-of-the-art techniques may enhance the accuracy and robustness of the face detection component. Future research efforts could focus on refining the face detector to better handle challenging conditions such as occlusions, varying lighting conditions, and complex facial appearances. By addressing these limitations, the overall performance and reliability of the proposed facial analytics system can be further enhanced.

VII. DISCUSSION

1) Enhanced Information Integration: Event cameras capture visual information differently from frame-based cameras, capturing sparse and timestamped events. Multi-task learning can utilize this unique data characteristic to train multiple facial analytics tasks simultaneously, such as head pose estimation, eye gaze estimation, and facial occlusion detection. This integration of information improves performance

across tasks by learning meaningful features from sparse event data.

2) Improved Robustness and Generalization: Multi-task learning enhances the robustness and generalization of facial analytics for event cameras. Jointly learning multiple tasks allows the model to capture underlying patterns and dependencies, resulting in a comprehensive representation of facial features. This improves adaptability to various scenarios, lighting conditions, and facial appearances, leading to better performance in real-world applications.
3) Leaky Integration and ROI Thresholding: In order to counteract global head motion, we present an innovative event integration method that effectively manages both short and long-term temporal dependencies. The proposed algorithmic technique combines leaky integration and ROI motion thresholding to mitigate interference from global subject motions in event stream data. By utilizing a leaky time surface and ROI inference triggering, accurate facial analysis is achieved. The approach optimizes facial analytics by selectively emphasizing events occurring within the face region, resulting in improved efficiency and accuracy.
4) Synthetic Data Generation: The research used V2E to simulate near-infrared (NIR) images to events, creating a large, high-quality synthetic dataset. Synthetic data generation is essential in computer vision research to overcome limitations of real-world data and develop diverse training datasets.
5) State-of-the-Art Results: The research achieved state-of-the-art results when evaluating the multitask network on the simulated event dataset. This demonstrates the effectiveness of the proposed approach, surpassing existing methods in facial analytics.
6) Comparable Results on Real-Event Data: The multitask network also achieved comparable results when tested on locally acquired event data. This indicates its ability to generalize to new scenarios, essential for practical applications of facial analytics.
7) Performance on BIWI Event Simulated Dataset: The trained model's performance on the BIWI event simulated dataset was not state-of-the-art. However, fine-tuning or retraining the model on this dataset could potentially improve its performance.
8) Exceptional Performance in Real-World Scenarios: The trained model was tested on real-world scenarios, where subjects were recorded using neuromorphic event cameras while driving. The results demonstrated exceptional performance, further validating the proposed approach's potential for driver monitoring systems and facial analytics applications.

This research demonstrates the potential of event-based facial analytics and opens up new avenues for research and development in the field. The proposed approach is highly effective and can be further improved with additional research, making it a promising direction for future work.

VIII. CONCLUSION

Event cameras offer several significant advantages over conventional cameras, making them particularly suited to the driver monitoring systems [8]. This paper presents a proof of concept and a baseline study on multi-task face analytics systems based on neural networks for such applications. The suggested model correctly analyzes occlusions, head pose, and estimates eye gaze. Moreover, a novel method to overcome short- and long-term information is presented. Event cameras react to motion. Thus, it only responds to changes or movements in the scene. Our method allows the multi-task network to also react to driver motion or changes only in the driver state, minimizing unnecessary computation. Thus, the rate of inference is linked to the speed of driver movement. The capabilities of event cameras are not only limited to the features presented in this paper and can enhance driver or human/driver monitoring systems beyond the limits of conventional cameras due to their high temporal resolution, low latency, high dynamic range, and low power consumption.

Future work will include collecting a larger and more diverse real-world event dataset that better reflects the variability and complexity of head poses, lighting conditions, occlusions, and other factors encountered in actual driver monitoring scenarios. Secondly, exploring different model architectures or modifications to the existing model to improve its accuracy in head pose estimation. This could include incorporating additional neural network layers, incorporating attention mechanisms, or leveraging other advanced techniques such as RCNN or CNN to capture temporal dependencies or spatial features more effectively. Moreover, investigating the potential of fusing event camera data with other sensor modalities, such as RGB cameras, depth sensors, or other physiological sensors, to improve the accuracy and robustness of the driver monitoring system. Multi-modal fusion approaches can leverage complementary information from different sensors to enhance the overall performance of the system.

REFERENCES

[1] Morris Bart. *Driver Error Causes Collisions*. Accessed: Mar. 3, 2023. [Online]. Available: https://www.morrisbart.com/blog/how-driver-error-causes-collisions/

[2] Lütkebohle. *BWorld Robot Control Software*. [Online]. Available: https://spyro-soft.com/blog/driver-monitoring-systems-to-become-mandatory-under-new-eu-and-us-road-safety-regulations

[3] C. Ryan, F. Murphy, and M. Mullins, "End-to-end autonomous driving risk analysis: A behavioural anomaly detection approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1650–1662, Mar. 2021.

[4] C. Ryan, "Emerging autonomous vehicle risks: The role of telematics and machine learning based risk assessment," Ph.D. thesis, Univ. Limerick, Limerick, Ireland, 2020.

[5] A. Koesdwiady, R. Soua, F. Karray, and M. S. Kamel, "Recent trends in driver safety monitoring systems: State of the art and challenges," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 4550–4563, Jun. 2017.

[6] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022, doi: 10.1109/TPAMI.2020.3008413.

[7] G. Lenz, S.-H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Frontiers Neurosci.*, vol. 14, p. 587, Jul. 2020.

[8] C. Ryan, B. O'Sullivan, A. Elrasad, A. Cahill, J. Lemley, P. Kielty, C. Posch, and E. Perot, "Real-time face & eye tracking and blink detection using event cameras," *Neural Netw.*, vol. 141, pp. 87–97, Sep. 2021.

[9] A. N. Angelopoulos, J. N. P. Martel, A. P. S. Kohli, J. Conradt, and G. Wetzstein, "Event based, near eye gaze tracking beyond 10,000 Hz," 2020, *arXiv:2004.03577*.

[10] M. A. Farooq, P. Corcoran, C. Rotariu, and W. Shariff, "Object detection in thermal spectrum for advanced driver-assistance systems (ADAS)," *IEEE Access*, vol. 9, pp. 156465–156481, 2021, doi: 10.1109/ACCESS.2021.3129150.

[11] 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.

[12] 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 Integr. Intell. Syst. (MFI)*, Sep. 2022, pp. 1–7.

[13] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 49–56, May 2016.

[14] 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.

[15] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "Anall-in-one convolutional neural network for face analysis," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 17–24.

[16] D. Yang, X. Li, X. Dai, R. Zhang, L. Qi, W. Zhang, and Z. Jiang, "All in one network for driver attention monitoring," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2258–2262.

[17] 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. Workshops (CVPRW)*, Jun. 2021, pp. 1312–1321.

[18] 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.

[19] W. Shariff, M. A. Farooq, J. Lemley, and P. Corcoran, "Event-based YOLO object detection: Proof of concept for forward perception system," in *Proc. 15th Int. Conf. Mach. Vis. (ICMV)*, vol. 12701. Bellingham, WA, USA: SPIE, Jun. 2023, pp. 74–80.

[20] 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.

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

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

[23] M. S. Dilmaghani, W. Shariff, C. Ryan, J. Lemley, and P. Corcoran, "Control and evaluation of event cameras output sharpness via bias," in *Proc. 15th Int. Conf. Mach. Vis. (ICMV)*, vol. 12701. Bellingham, WA, USA: SPIE, Jun. 2023, pp. 455–462.

[24] P. Liu, G. Chen, Z. Li, D. Clarke, Z. Liu, R. Zhang, and A. Knoll, "NeuroDFD: Towards efficient driver face detection with neuromorphic vision sensor," in *Proc. Int. Conf. Adv. Robot. Mechatronics (ICARM)*, Jul. 2022, pp. 268–273.

[25] G. Moreira, A. Graça, B. Silva, P. Martins, and J. Batista, "Neuromorphic event-based face identity recognition," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 922–929.

[26] Y. Feng, N. Goulding-Hotta, A. Khan, H. Reyserhove, and Y. Zhu, "Real-time gaze tracking with event-driven eye segmentation," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*, Mar. 2022, pp. 399–408.

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

[28] G. Chen, F. Wang, W. Li, L. Hong, J. Conradt, J. Chen, Z. Zhang, Y. Lu, and A. Knoll, "NeuroIV: Neuromorphic vision meets intelligent vehicle towards safe driving with a new database and baseline evaluations," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 1171–1183, Feb. 2022.

[29] A. Elrasad, C. Ryan, R. Blythman, J. Lemley, and B. O'Sullivan, "Event detector and method of generating textural image based on event count decay factor and net polarity," U.S. Patent 11 270 137, Mar. 8, 2022.

[30] C. Ryan, R. Blythman, J. Lemley, A. Elrasad, and B. O'Sullivan, "Object detection for event cameras," U.S. Patent 11,301,702, Apr. 12, 2022.

[31] G. Chen, F. Wang, X. Yuan, Z. Li, Z. Liang, and A. Knoll, "NeuroBiometric: An eye blink based biometric authentication system using an event-based neuromorphic vision sensor," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 1, pp. 206–218, Jan. 2021.

[32] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1871–1880.

[33] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.

[34] G. Hu, L. Liu, Y. Yuan, Z. Yu, Y. Hua, Z. Zhang, F. Shen, L. Shao, T. Hospedales, N. Robertson, and Y. Yang, "Deep multi-task learning to recognise subtle facial expressions of mental states," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 103–119.

[35] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[36] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, NJ, USA, Jun. 2020, pp. 3586–3595.

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

[38] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, "How to train your event camera neural network," 2020, *arXiv:2003.09078*.

[39] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, "Reducing the sim-to-real gap for event cameras," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 534–549.

[40] A. Zanardi, A. J. Aumiller, J. Zilly, A. Censi, and E. Frazzoli, "Cross-modal learning filters for RGB-neuromorphic wormhole learning," in *Proc. Robot., Sci. Syst. XV*, Jun. 2019, p. P45.

[41] S. Afshar, T. J. Hamilton, J. Tapson, A. van Schaik, and G. Cohen, "Investigation of event-based surfaces for high-speed detection, unsupervised feature extraction, and object recognition," *Frontiers Neurosci.*, vol. 12, p. 1047, Jan. 2019.

[42] B. Ramesh, S. Zhang, Z. W. Lee, Z. Gao, G. Orchard, and C. Xiang, "Long-term object tracking with a moving event camera," in *Proc. BMVC*, 2018, p. 241.

[43] B. Ramesh, S. Zhang, H. Yang, A. Ussa, M. Ong, G. Orchard, and C. Xiang, "E-TLD: Event-based framework for dynamic object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3996–4006, Oct. 2021.

[44] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1087–1096.

[45] B. Huang, R. Chen, W. Xu, and Q. Zhou, "Improving head pose estimation using two-stage ensembles with top-k regression," *Image Vis. Comput.*, vol. 93, Jan. 2020, Art. no. 103827.

[46] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019.

[47] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "FDN: Feature decoupling network for head pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12789–12796.

[48] 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, pp. 437–458, Feb. 2013.

[49] X. Zheng, Y. Liu, Y. Lu, T. Hua, T. Pan, W. Zhang, D. Tao, and L. Wang, "Deep learning for event-based vision: A comprehensive survey and benchmarks," 2023, *arXiv:2302.08890*.

[50] C. Boretti, P. Bich, F. Pareschi, L. Prono, R. Rovatti, and G. Setti, "PEDRo: An event-based dataset for person detection in robotics," in *Proc. CVPR*, Jun. 2023, pp. 4064–4069.

[51] Z. Zhang, K. Chai, H. Yu, R. Majaj, F. Walsh, E. Wang, U. Mahbub, H. Siegelmann, D. Kim, and T. Rahman, ''Neuromophic high-frequency 3D dancing pose estimation in dynamic environment,'' in *Proc. CVPR Workshop*, 2023.

[52] L. Berlincioni, L. Cultrera, C. Albisani, L. Cresti, A. Leonardo, S. Picchioni, F. Becattini, and A. D. Bimbo, ''Neuromorphic event-based facial expression recognition,'' 2023, *arXiv:2304.06351*.

[53] Y. Wang, W. Zhou, and J. Zhou, ''2DHeadPose: A simple and effective annotation method for the head pose in RGB images and its dataset,'' *Neural Netw.*, vol. 160, pp. 50–62, Mar. 2023.

[54] W. Deng and R. Wu, ''Real-time driver-drowsiness detection system using facial features,'' *IEEE Access*, vol. 7, pp. 118727–118738, 2019, doi: 10.1109/ACCESS.2019.2936663.

[55] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, ''YOLO-face: A real-time face detector,'' *Vis. Comput.*, vol. 37, no. 4, pp. 805–813, Apr. 2021.

[56] P. Chakraborty, S. Ahmed, M. A. Yousuf, A. Azad, S. A. Alyami, and M. A. Moni, ''A human–robot interaction system calculating visual focus of human's attention level,'' *IEEE Access*, vol. 9, pp. 93409–93421, 2021.

[57] S. Ahmad, G. Scarpellini, P. Morerio, and A. D. Bue, ''Event-driven re-ID: A new benchmark and method towards privacy-preserving person re-identification,'' in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa, HI, USA, Jan. 2022, pp. 459–468, doi: 10.1109/WACVW54805.2022.00052.

CIAN RYAN received the M.Sc. and Ph.D. degrees in computational finance from the University of Limerick, in 2015 and 2020, respectively. He is currently a Staff Machine Learning Engineer with Xperi Inc., Galway. His work focuses on machine learning and deep learning methods applied to computer vision.

AMR ELRASAD received the B.Sc. degree in electrical engineering and M.Sc. degree in engineering mathematics from Alexandria University, Egypt, in 2004 and 2010, respectively, and the Ph.D. degree in computer science from KAUST in 2016. He is currently a staff Research and Development Machine Learning Engineer with Xperi Inc. His research interests include health monitoring for indoor and in-cabin applications.

WASEEM SHARIFF received the B.E. degree in computer science from the Nagarjuna College of Engineering and Technology (NCET), in 2019, and the M.Sc. degree in computer science, specializing in artificial intelligence from the National University of Ireland Galway (NUIG), in 2020, where he is currently pursuing the Ph.D. degree under the IRC Employment Ph.D. Program. He is a Research and Development Engineer with Xperi Inc., Galway. His research interest includes machine learning for computer vision applications, with a particular emphasis on automotive driver monitoring applications.

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

PAUL KIELTY received the B.E. degree in electronic and computer engineering from the University of Galway, in 2021. He is currently pursuing the Ph.D. degree with the University of Galway and the ADAPT SFI Research Centre. His research interest includes deep learning methods with neuromorphic vision, with a particular interest in driver monitoring tasks.

PATRICK HURNEY received the B.Eng. degree in electronic and computer engineering from the School of Engineering, University of Galway, in 2009, and the Ph.D. degree in electronic engineering from the University of Galway, with a focus on real-time detection of pedestrians in infrared images, in 2015. He is currently a Machine Learning Research Engineer with Xperi Inc., Galway, Ireland. His research interests include machine learning with a focus on the automotive safety of both vulnerable road users and in-vehicle occupants and the real-time performance of computer vision-based machine learning algorithms on EDGE devices.

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