FatigueView

Citation for published version:

Digital Object Identifier (DOI):
10.1109/TITS.2022.3216017

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Peer reviewed version

Published In:
IEEE Transactions on Intelligent Transportation Systems

Publisher Rights Statement:
© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

General rights
Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
FatigueView: A Multi-Camera Video Dataset for Vision-Based Drowsiness Detection

Cong Yang, Zhenyu Yang, Weiyu Li, and John See, Senior Member, IEEE

Abstract—Although vision-based drowsiness detection approaches have achieved great success on empirically organized datasets, it remains far from being satisfactory for deployment in practice. One crucial issue lies in the scarcity and lack of datasets that represent the actual challenges in real-world applications, e.g., tremendous variation and aggregation of visual signs, challenges brought on by different camera positions and camera types. To promote research in this field, we introduce a new large-scale dataset, FatigueView, that is collected by both RGB and infrared (IR) cameras from five different positions. It contains real sleepy driving videos and various visual signs of drowsiness from subtle to obvious, e.g., with 17,403 different yawning sets totaling more than 124 million frames, far more than recent actively used datasets. We also provide hierarchical annotations for each video, ranging from spatial face landmarks and visual signs to temporal drowsiness locations and levels to meet different research requirements. We structurally evaluate representative methods to build viable baselines. With FatigueView, we would like to encourage the community to adapt computer vision models to address practical real-world concerns, particularly the challenges posed by this dataset.

Index Terms—Drowsiness detection, driver monitoring system, intelligent cockpit, autonomous vehicles.

I. INTRODUCTION

Drowsiness detection is necessary for various application domains, including driver monitoring system (DMS) [1], smart-office [2] and smart-home [3], [4], etc. For instance, timely detection of drowsiness is a key function in DMS to avoid drowsy driving and potential accidents [5]. Following the convention in [1], [6], [7], [8], and [9], an intuitive purpose of vision-based drowsiness detection is to recognize visual signs (see Fig. 1) on human face and body from camera-captured image sequences, thereby to determine the occurrence and stage/level of drowsiness. Though vision-based drowsiness detection has been extensively studied in the past decades [10], [11], its usability remains limited in accurate and robust detection scenarios [12], [13]. One crucial issue is the scarcity of public datasets that represent the challenges in real-world applications.

Firstly, the expressions of drowsiness are vary greatly, ranging from very subtle cues to the obvious. In most cases, they involved the interaction of a few visual signs. However, most existing datasets only focus on independent and explicit visual cues and it is too restrictive for practical usage. For example, a yawning is normally accompanied by hand on mouth gesture and a prolonged eye closure (Fig. 2 (a)). In such a case, traditional yawning-based drowsiness detection approaches [14], [15] cannot be properly applied due to the initial occlusion. Secondly, camera positions are not fixed at a single position in practice but existing datasets and their corresponding algorithms are mostly based on fully frontal face views [1]. For instance, a camera could be placed at any of these six positions (i.e., steering column, A-Pillar, instrument cluster, dashboard, center console and rear-view mirror) in DMS scenarios and the challenges introduced by these positions are significantly different. In Fig. 2 (b), when the camera is placed on the A-Pillar (top-left), most eye blinking approaches are more likely to fail due to the occlusion from a likely glass frame. It is also difficult to tell the eye state of the driver in Fig. 2 (c) between closure and looking down. As a result, most approaches cannot be fully assessed since such practical challenges are barely covered by existing datasets. Thirdly, in some systems both RGB (Fig. 2 (d)) and IR cameras are employed to cope with the nighttime and poor/dimmed lighting conditions [16]. However, to our best knowledge, only a few datasets such as [1] and [17] contain both camera types. In such conditions, it is generally difficult to robustly detect signs of drowsiness, especially with subtle body and facial signs. Lastly, the annotations on existing datasets are only focused on extreme drowsiness and explicit signs [1], [18], which are inadequate to push the boundaries of research and applications. In practice, fine-grained annotations are as equally important for early detection of driver drowsiness.

In this paper, we introduce a large-scale video dataset, FatigueView, to promote research in this field. It contains...
both training and testing data, which are distinct in three aspects: 1) Practical. The training data is recorded with the guidance of FatigueTree, an open and structural collection of visual signs observed from real-world videos. The testing data contains two sets: (a) real sleepy driving videos of intercity coach and truck drivers and, (b) drowsiness videos from awake to fully fall asleep that are recorded in the simulated DMS scenario. Both training and testing data represent various visual signs (around 1,661,574 in number) ranging from subtle to obvious. 2) Diversity. Except real sleepy driving videos, each video in FatigueView is recorded by both RGB and IR cameras at five different positions. This contributes to various challenges introduced by different camera positions and types. 3) Richness. FatigueView provides a solid foundation to facilitate the study of drowsiness detection. This is built on the fact that FatigueView has rich annotations from coarse to fine (Elements, Sets and Events), both temporally and semantically, and thus providing more flexibility in modelling. A series of empirical studies are conducted using existing methods to reveal the challenges in practice. These aspects together provide a solid basis towards a reliable benchmark for assessment.

The main contributions are: (1) the creation of a large-scale video dataset for drowsiness detection, FatigueView, with distinguished features including practical, diversity and richness. For this, we introduce a systematic and general framework for drowsiness data collection and annotation. (2) We benchmark state-of-the-art algorithms on FatigueView, establishing a sizeable number of baselines for future algorithm development. Our evaluation and analysis demonstrate that a significant amount of research is still necessary before existing techniques can be reliably applied in practical applications that require robust drowsiness detection, particularly in its early stage of occurrence.

II. RELATED WORKS

We now present a concise survey of existing datasets and their collection strategies, followed by a snapshot of some vision-based drowsiness detection methods. For a more thorough treatment on drowsiness detection methods, recent compilations by Sparrow [9] and Ghoddoosian [18] offer sufficiently good reviews.

A. Drowsiness Datasets

Table I presents a short summary of drowsiness-related datasets in literature, which are publicly (or partly) available at the time of writing. In terms of camera positions and types, a large number of datasets recorded people’s frontal face with an RGB camera and hardly mimic real-world challenges. Among the public datasets, ZJU [47] and YawDD [48] are commonly used datasets for specific tasks of eye blinking and yawning detection, respectively. However, as discussed in Section I, such independent visual cues are still insufficient for a comprehensive study on driver drowsiness. For vision-based drowsiness detection, NTHU [1] is the most actively used one among the listed as it contains a large number of videos with a diverse set of participants and assessment benchmark methods. However, it has limitations in representing real-world challenges due to its scant camera positions and types. Compared to all existing datasets, the proposed FatigueView leads in both quantity and quality. Moreover, the hierarchical annotations of FatigueView offer desirable coverage of most related research tasks targeted by other datasets.

Considering collection strategies, around 65% datasets in Table I were collected by acting in a stopped car, lab and simulated environments. There is no guarantee that the data collected from acting would properly reflect the reality. Though recent published DMD [52] contains various real driving videos from car and simulated scenarios, it covers only three simple signs and not fully released yet. This is understandable since it is extremely dangerous to collect real sleepy driving events in practice. With the help of DMS volunteers, FatigueView contains real sleepy driving videos in-the-wild for testing. Our testing data also contains full period of driving from awake to real fall asleep in the simulated DMS environment. Better yet, our training data was collected by the guidance of FatigueTree, which can dramatically improve the data quality and reflection of reality.

B. Drowsiness Detection Methods

Though various modalities can be employed in detecting drowsiness patterns from body signals [57], [58] and driving actions [59], [60], recent works mainly focus on using vision-based approaches [10], [15], [20], [61], [62]. This is because compared to body-carrying sensors [57], [58], vision-based methods are touchless and naturally more convenient in real-world applications, popularly driven by the use of deep learning [14], [61], [63], [64].

1) Eye State: The closed eye state is a crucial element that has been considered. For instance, the Percentage of Eyelid Closure (PERCLOS) is used reliably to measure predicted drowsiness levels by calculating the percentage of duration of closed-eye state in a specific time interval [65], [66]. The eye status can be estimated based on the eye white part reflection [34] and the eye shape [25] while more recent works [37], [67] use Convolutional Neural Network (CNN) to classify eye status based on the cropped eye region. Alternatively, the distance between eyelids (DBE) [38] and eye aspect ratio (EAR) [68] can be computed to estimate eye status using outside and inner eye landmarks. The use of eye landmarks dates back to early attempts [36] of utilizing optical flow for eye status classification.

2) Yawning State: In addition to eye status, yawning is also a vital visual sign for drowsiness detection at an early
stage [15]. To ascertain a yawning state, a classic technique involved the measuring of both the rate and intensity of changes in the mouth area [15]. Meanwhile, some works attempt to model yawning using supervised [69] and unsupervised [70] learning techniques.

3) Fusion: As drowsiness signs obtained independently from the eye and mouth could be easily affected by lighting, illusion and occlusion, more comprehensive methods were introduced to improve detection robustness. For example, Bergasa et al. [71] fused PERCLOS, eye closure duration and fixed gaze while Ying et al. [72] tried to detect driver drowsiness based on a combination of eye and mouth gestures. More recently, end-to-end deep neural networks have been proposed to learn features from an array of different drowsiness signs [1], [14], [18], [61], [67]. Taking advantage of the new opportunities offered by FatigueView, we conduct a series of

### TABLE I

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Purpose</th>
<th>Camera and Position</th>
<th>People</th>
<th>Quantity</th>
<th>Environment</th>
</tr>
</thead>
</table>
| CCDWC [19] | Eye Blinking | 2 RGB, instrument cluster | - | 15 videos, 250-400 frames each | simulated, act | No *
| OUD [20] | Eye Status | 1 RGB, frontal face | 23 | 23 videos, 30 min each | simulated, real | No *
| RDD [14] | Drowsiness | 1 RGB, frontal face | 33 | 99 videos, 1-1.5 min each | act, car | No *
| MBG [21] | Drowsiness | 1 IR, instrument cluster | 23 | 30 real drive videos | car, real | No *
| DDD [22] | Eye Closeness | 1 IR, instrument cluster | 21 | around 100,000 images | car, act | No *
| SRV [23] | Eye Closeness | 1 RGB, frontal face | 319 | 2314 cropped eye images | lab, act | No *
| ADD [24] | Head Pose | 1 RGB, frontal face | 1 | - | lab, act | No *
| SBE [25] | Eye Blinking | 1 RGB, frontal face | - | 21 videos, 1 min each | lab, act | No *
| MHE [26] | Head/Eye Motion | 1 RGB, dashboard | 2 | - | car, act | No *
| SEU [27] | Drowsiness | 1 RGB, side-mounted | 20 | 80 driving posture images | car, act | No *
| AOD [28] | Drowsiness | 1 RGB, frontal face | 21 | 42 videos, 60 min each | simulated, act | No *
| MCSO [29] | Drowsiness | 2 RGB, frontal face | 15 | 45 videos, 10-60 min each | simulated, act | No *
| CDL [30] | Drowsiness | 1 RGB, frontal face | 49 | 49 videos, 120 min each | simulated, real+act | No *
| FDU [31] | Drowsiness | 1 RGB, frontal face | 20 | 20 videos, 120 min each | car+simulated, - | No *
| HCV [32] | Eye Status | 1 RGB + 1 IR | - | - | car, act | No *
| AOF [33] | Yawning | 1 RGB, frontal face | 12 | 123 yawning videos, 6.28 sec each | simulated, real | No *
| MOD [11] | Eye Blinking | 1 RGB, frontal face | 6 | 18 videos, 12 min each | simulated, act | No *
| DDF [34] | Eye Status | 1 RGB, frontal face | - | 4 videos, 433-2717 frames each | - | act | No *
| TES [35] | Eye Status | Frontal face | - | 150 images | car, act | No *
| EBB [36] | Eye Blinking | 1 RGB | 3 | 10 videos, 20,000 frames in total | lab, act | No *
| IRF [37] | Eye Status | 1 IR, frontal face | 20 | 160 videos | lab, act | No *
| HDB [38] | Eye Status | 1 RGB, frontal face | - | 10 videos, 7020-8572 frames each | lab, act | No *
| DAS [39] | Eye/Mouth Status | 1 IR, side-mounted | 5 | - | car, act | No *
| DHD [40] | Drowsiness | Frontal face | - | 4 videos, 28,000 frames in total | car-, - | No *
| DFF [41] | Eye Status | 1 RGB, frontal face | - | - | lab, act | No *
| HPA [42] | Eye Blinking | 1 RGB + 1 IR, frontal face | - | - | car, act | No *
| DSD [43] | Drowsiness | 2 RGB + 2 IR, instru. cluster | 14 | - | car, act | No *
| VAO [44] | Drowsiness | 1 RGB, frontal face | 5 | 15 videos, 5 min each | car, act | No *
| MDC [45] | Drowsiness | 1 Kinect, frontal face | 35 | 105 videos, 10 min each | lab, act | No *
| EyeblinkKin [45] | Eye Blinking | 1 RGB, frontal face | 4 | 8 videos | lab, act | No *
| DriveAHead [46] | Head Pose | 1*(IR+Depth), frontal face | 20 | 21 videos, 30 min each | car, real | No *
| ZIJU [47] | Eye Blinking | 1 RGB, frontal face | 20 | 80 eye blinking videos | lab, act | Yes *
| NTHU [1] | Drowsiness | 1 RGB + 1 IR, A-Pillar | 36 | 360 videos, 1-1.5 min each | simulated, act | Yes *
| YawDD [48] | Yawning | 2 RGB, dashboard & RVM | 107 | 342 videos, 15-40 sec each | car+act=real | Yes *
| DROZY [17] | Drowsiness | 1 RGB + 1 IR, frontal face | 14 | 14 videos, 10 min each | lab, real | Yes *
| CEW [49] | Eye Closeness | Colour images | - | 2423 cropped eye images | - | - | Yes *
| RT-BENE [50] | Eye Blinking | 1 Kinect, frontal face | 15 | 243,714 frames | lab, act | Yes *
| 300-VW [51] | Facial ldmk | 1 RGB, frontal face | - | 50 videos | natural, - | Yes *
| RLDD [18] | Drowsiness | 1 RGB, frontal face | 12 | 180 videos, 10 min each | lab, real | Yes *
| DMD [52] | Comprehensive | 3*(RGB+IR+Depth), 3 locat. | 37 | 41 hours | car+simulated,real+act | Yes *
| DriveAct [53] | Behaviour | 5 IR + 1 Kinect, 6 locat. | 12 | 12 hours | simulated, act | Yes *
| Pandora [54] | Driver Pose | 1 RGB + 1 Kinect, frontal face | 22 | 110 videos | lab, act | Yes *
| DD-Pose [55] | Head Pose | 1*(RGB+Depth), frontal+side | 27 | 330 frames | car, real | Yes *
| CMU-PIE [56] | Head Pose | 1 RGB, frontal face | 72 | 1503 images | lab+car, act+real | Yes *

**FatigueView** | EMBHD | 5 RGB + 5 IR, 5 locations | 95 | 124.44M frames, 1,384 hours | car+simulated, real+act | Yes *

(EMBHD: Eye/Mouth/Head/Body and Drowsiness. RVW: Rear-View Mirror. -: Unclear. FBH: Face/Body/Hand.)
III. FATIGUEVIEW

Fig. 3 presents an overview of the FatigueView data collection framework, which is specifically guided by FatigueTree. Training data are both long and short acted videos from participants while the testing data consists of driving videos in real and simulated DMS scenarios. The framework is motivated by the fact that the training data should cover as many visual signs as possible while the testing data should allow real scenarios to be predicted.

A. FatigueTree

As presented in Fig. 3 (a) and (b), both real video collection and FatigueTree construction are carried out semi-automatically in human-in-the-loop fashion. To initialize FatigueTree, we manually search and download videos from various websites (e.g., DMS feedbacks, film, television and YouTube, etc.) using drowsiness-related keywords in 6 languages (Arabic, Chinese, English, French and Spanish). With the assistance of existing face and eye landmark detectors [73], the videos are then analysed by human experts to construct a FatigueTree by summarizing the observed visual signs and their characters until the tree stabilizes. Specifically, the taxonomy of FatigueTree is motivated by (1) our statistics of real drowsiness videos according to body parts (including dressing and characteristics, e.g., cross-eye), duration, degree of expression, sensor location (and type), and (2) industrial practice of auto manufacturers. For instance, NIO’s ES8 model mainly focus on yawning and prolonged eye closure features. Changan’s UNI-T model not only detect these two features, but also frequent blinking. Accordingly, a drowsiness event could be composed of movements (or status) observed at the mouth (M), eye (E), upper body (B) or combinations of them.

Taking the “Eye branch” as an example, its drowsiness-related sub-branches include different eye statuses (see Fig. 4), occlusions by glasses and hats, etc. For the mouth, most features are related to yawning and few of them are slight mouth opening along with head and eye features. Thus, visual signs in the “mouth branch” include mouth occluded by hand, opening mouth slowly, long-time-yawning, short-time-yawning, slight yawning, big and mouth yawning (long time and short time), etc. In terms of the body, the most distinct feature is stretching. Beyond that, there are more complicated combinations from hand, head, and face. For instance, holding head with hands, itching face and neck, rubbing eyes, head tilting, upturning, nodding, and downturning, etc. All these features have correlated sub-branches and sample drowsiness videos in the FatigueTree. Thus, FatigueTree not only caters towards real-world challenges in academic research, but also to facilitate standardization of taxonomies from industrial partners.

In practice, FatigueTree is an online system with a taxonomic visualization (tree diagram) and submission functions. If the drowsiness type is not found in the tree, users can add new relevant branches to the tree diagram and upload the real video (including the description) via the submission page. We also utilise several procedures to avoid ambiguity and ensure reliability when inserting video clips. Specifically, during the human-in-the-loop insertion process, our guiding
system dynamically asks the submitter regarding the subjects in the clip based on the tree nodes. In each step, some sample videos from the closed nodes are also shown for comparison. Once a certain visual sign is determined in a clip, the administrator will check the clip quality before internal merging. Since FatigueTree was employed by some industrial stakeholders and research labs, their communities meet regularly to check, discuss and arbitrate on the newly added branches and videos to ensure reliability. During the meeting, we also use inter coder reliability tests to measure the consistency of their decisions. Currently, there are 173 visual signs in FatigueTree categorized based on pure “eyes”, “hands”, “head”, “mouth” and their 14 combinations. Particularly, these combinations are existed and actively occurred in real drowsiness events. For instance, “eyes, hands”, “eyes, hands, head”, “eyes, head”, “eyes, mouth”, “eyes, mouth, hands”, “eyes, mouth, head”, “hands, head”, “mouth, hands”, “mouth, head”, etc. For each combination, the order of elements is based on their visual explicitness. Thus, some sub-branches may have the same elements but different orders. To improve the efficiency of FatigueTree management in practice, looking for a combination is normally a back-and-forth task: decomposing a drowsiness event to sets, or verifying a possible combination in real drowsiness videos.

The rationale behind FatigueTree is to collect various real drowsiness visual signs in a natural way while managing them structurally. Besides, FatigueTree can also effectively avoid duplicated drowsiness types so as to improve data collection efficiency. To aid the participants, we generate various long and short videos (2-10 minutes) in a controlled way by combing through the real videos in leaf nodes and also non-drowsiness videos to cover different types and duration of drowsiness. These videos were used for guiding the participants’ acting via the tablet (Fig. 3 (c)). To ensure the applicability and coverage of these viewing videos, we not only strictly follow the temporal stages of different drowsiness types, but also keep reasonable flexibility on the combination of visual signs. For instance, eye blinking is randomly accompanied by yawning, rubbing eyes, wearing glasses, etc.

B. Environment

For safety reasons, only simple visual signs (like yawning) could be collected during real driving. It is extremely dangerous to collect real sleepy driving events in practice, unless in a stopped car [42], [44], [48]. Different from that, our simulated environment is essentially similar to driving scenario in a stopped car [42], [44], [48]. Different from that, our simulated environment is essentially similar to driving scenario in a stopped car [42], [44], [48]. Different from that, our simulated environment is essentially similar to driving scenario. In contrast, the testing data are real drowsiness videos that can effectively bring richer and more variability in data. To improve the concentration of participants in the driving scene, three strategies were introduced during capturing: (1) Benefit from the featured driving simulator, precisely recreated territories, large screen, and our physical equipment, participants had high immersion in the simulated environment.

Each participant is asked to sit and act out the drowsiness state with the guidance of the viewing videos from the tablet (Fig. 3 (c)). Each participant is also given the freedom to perform their act based on their own understanding and personal habit of a certain drowsiness type. Such flexibility can effectively bring richer and more variability in data. In contrast, the testing data are real drowsiness videos that were captured in a simulated DMS scenario (Fig. 3 (d)). To ensure a realistic setup, a participant is equipped with a simulated driving wheel with pedals (Thrustmaster T150RS). During recording, the participant sits on a fixed but tunable chair and plays a driving game (Euro Truck Simulator 2) with soft ambient background music.
Section IV-C). The viewing videos in this phase are also collected in this phase for the purpose of distraction inaging videos to act drowsiness-related visual signs. Some easily the set level recording, each participant is guided by the view-coverage visual signs in FatigueTree. In phase one, which is FatigueView are similar to reality.

Training data are recorded in two phases to ensure the coverage visual signs in FatigueTree. In phase one, which is the set level recording, each participant is guided by the viewing videos to act drowsiness-related visual signs. Some easily confused non-drowsiness signs (e.g. yawning and sneezing) are also collected in this phase for the purpose of distraction (see Section IV-C). The viewing videos in this phase are short and independent, abstracted from FatigueTree leaves of high occurrence rate. The main purpose of phase one is to ensure the explicit visual signs of different participants can be captured. Moreover, participants could gradually familiarize themselves with the recording environment and the basic actions after the first phase. Phase two is the event level recording, whereby participants are inspired by the longer viewing videos to act out different drowsiness events ranging from wakefulness to sleep.

There are several crucial reasons to use acted videos: (1) real sleepy driving videos are limited in quantity (very few people and extremely dangerous to collect) and quality (mostly with only one camera) in practice. (2) It is expensive to locate and annotation drowsiness events from in-the-wild driving videos. Differently, our strategy can reduce the cost of labelling and preserve data quantity, diversity and quality. Specifically, acting in FatigueView is supervised by FatigueTree, which has several benefits that close the gap towards reality: (1) Various realistic signs (both subtle and obvious) are all structurally preserved in FatigueTree and then properly reflected in the training data. (2) 3 additional strategies to ensure consistency in acting: (a) training before acting, (b) acting from sign to event levels, (c) acting with freedom of improvisation.

For testing data collection (Test-Sim), different from those in the training data, were recruited at random yet ensuring balance in a variety of dressing accessories like glasses. To balance the drowsiness and normal states, each participant performed two rounds of video recordings: day time between 12:00 am–15:00 pm and night between 0:00 am–3:00 am. For more effective recording, each participant was deprived of sleep and caffeine for 18 hours before taking the driving task (for the night round). Participants were informed of the possible risks of this experiment before agreeing to participate. To cover more realistic events, participants are free to eat and drink during driving. We also randomly create over-and under-exposed scenes by moving (or covering) the IR pods and overhead lamps to simulate influences from lights. This is because sunlight intensity, coverage, and angle could potentially influence the performance of drowsiness detection algorithms. For instance, extreme sunlight may cause squinting and moving lamps can mimic various light conditions.

D. Annotation

Inspired by FineGym [76], we provide hierarchically based annotations for our data. As shown in Fig. 3 (e), the annotations are organized according to a three-level hierarchy: Elements, Sets and Events. Elements, at the finest level of the hierarchy, refer to essential features on each frame, including face and hand bounding boxes, facial landmarks (68 points), head pose (yaw, pitch, row), eye status (open, closure, squinting), eye landmarks (17 points), skeleton (upper body). These are features that are commonly used for detecting drowsiness visual signs. An example in Fig. 6 (b) shows that the 17 eye landmarks include 8 for eyelids, 8 for the iris and one pupil center point. We use a heuristic process to annotate frames in the element level. For instance, we first select a small set of frames for skeleton, head pose, face landmark pre-labelling using the adapted HigherHRNet [77], MLA [78] and AWing [73], respectively. The frames are then submitted to our labelling system for manual correction. This is repeated until a stable accuracy is achieved. Finally, we visualize the annotations in fragments to cross-check the data quality.

Sets are short videos, describing drowsiness-related visual signs like yawning, nodding, stretching, three types of blinking, prolonged eye closure, touching face, heavy-laden eyes, squeezing eyelids, shaking head quickly and non-drowsiness states like drinking and eating (see Fig. 7). Theoretically, a set
can be detected and localized based on elements. Meanwhile, at the coarsest granularity are event categories which are composed of sets. Specifically, if a video segment contains a high density of drowsiness-related visual signs (sets), the whole segment can be annotated as a drowsy event with statistics of visual signs (e.g. frequency and duration, etc.). The number and type of visual signs reflect the level of a drowsiness event as well as the sensitivity of a drowsiness detection system. FatigueView provides two levels of temporal annotations in the training and testing data, namely locations of all drowsiness events in a video and locations of all drowsiness-related visual signs, or sets in an event. With this structure, researchers can flexibly redefine the labels of stages based on their own scenarios. In terms of annotation team, our annotators were divided into three teams: (1) initial labelling team (60%, with more than 8 hours' training), (2) quality checking team (30%, engineers with DMS development experience), and (3) final confirmation team (10%, researchers and engineers with real experience in drowsiness detection systems). This workflow ensures that each annotation is double checked by experienced engineers and finally confirmed by experts.

E. Real Driving Videos

Our testing data also contains real driving videos (Test-Real) that are selected by experts from two implemented DMS systems in intercity coaches and trucks (both day and nights). The original videos were recorded by surveillance cameras located in steering column (or dashboard). In total, It includes 833 videos (3-360 seconds each, 640 × 480 and 25 fps) from 15 volunteers. These videos cover normal driving, distraction, and different levels of drowsiness. More importantly, it contains extremely dangerous distraction and drowsiness events. As shown in Fig. 8, videos are labelled by considering both visual signs and driving speed (>50 km/h). In other words, these real events could potentially result in traffic accidents. Due to privacy concerns, meta data (name, face, license, etc.) of drivers are confidential. Thus, except visual cue-related eyes and mouth, their faces are blurred with mosaic based on 68 facial landmarks. To facilitate different research tasks, hierarchical annotations on this set are all provided based on original frames.

IV. Properties

FatigueView\textsuperscript{1} offers some noteworthy properties: a complex mix of various visual signs, distinct drowsiness levels and real-world challenges. In total 1384 hours, the training and testing videos clocked up to around 655 (around 59 million frames) and 729 (around 66 million frames) hours, respectively.

\textsuperscript{1}The FatigueView, FatigueTree, and the codes of experiments are available at: https://fatigueview.github.io/ and https://github.com/FatigueView/.

TABLE II

<table>
<thead>
<tr>
<th>Visual Signs</th>
<th>Yawning</th>
<th>EC</th>
<th>Blinking</th>
<th>Nodding</th>
<th>Stretching</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>10590</td>
<td>22260</td>
<td>905360</td>
<td>8250</td>
<td>920</td>
<td>947380</td>
</tr>
<tr>
<td>Test-Sim</td>
<td>6550</td>
<td>8730</td>
<td>683470</td>
<td>150</td>
<td>310</td>
<td>701210</td>
</tr>
<tr>
<td>Test-Real</td>
<td>263</td>
<td>49</td>
<td>16115</td>
<td>1</td>
<td>29</td>
<td>16457</td>
</tr>
</tbody>
</table>

Average clip length: Train (52 s), Test-Sim (62 m), Test-Real (40 s)

A. Various Visual Signs

Table II presents the number of explicit visual signs and participants in the training and testing data. We clearly find that FatigueView surpasses existing datasets in both quality and quantity. In comparison to existing datasets that concentrate on a maximum of two visual signs (see Table I), FatigueView contains set level annotations on five explicit visual signs. Besides, more subtle visual signs can be further explored based on annotations at the element level. For instance, all 95 participants have both features of drowsiness and non-drowsiness events. Quantity-wise, there are 17,403 different yawning (video) sets in FatigueView, which is far more than the popular YawDD [48] with only 342 yawning sets in total. In addition, FatigueView contains 1,606,945 eye blinking sets from different camera types and locations, which is significantly more compared to another recent dataset RT-BENE [50] with 10,000 samples and only one camera located at the frontal face position. We also provide set level annotations for Nodding and Stretching, two traits that are closely related to drowsiness but not covered in existing datasets.

Building upon the large-scale annotations provided, we provide further insights based on the properties of different visual signs. For example, in testing data we can find that blinking occurs most often among all visual signs, followed by prolonged eye closure and yawning. Thus, it is reasonable to put more effort into optimizing eye status-related algorithms. Fig. 9 presents the statistical distribution of the duration of four visual signs on their training data. We observe that most yawning and blinking sets lasts around 2-7 and 0.2-0.6 seconds.

Fig. 8. Real sleepy driving videos in FatigueView. For privacy, driver faces are blurred with mosaic, except eyes and month.

Fig. 9. Distribution of duration (in seconds) of visual signs.
Alert occurs at all intervals while levels at each 10-minute time interval (Fig. 10), we find that the mapping rules. Based on the statistics of five drowsiness considerable amount of Signs is already close to dangerous driving, it is practical to put fully awake and frequency yawning; Signs indicate heavy eyes with signs of sleepiness like prolonged squinting, heavy-laden eyes, high occasionally rubbing eyes; Sleepy means eye blinking with prolonged eye closure and nodding. The other intervals are Normal being fully awake and Sleep with long-term eye closure [80], [81].

FatigueView can be adapted to different indicators with the mapping rules. Based on the statistics of five drowsiness levels at each 10-minute time interval (Fig. 10), we find that Alert occurs at all intervals while Sleepy is accompanied by a considerable amount of Signs in most intervals. Since Sleepy is already close to dangerous driving, it is practical to put more effort towards detecting Signs event. Such insights are valuable for researchers and developers to draw observations for practical deployment.

C. Various Challenges

With the proposed collection strategies, FatigueView can properly preserve various challenges brought about by different camera positions, camera types and events. For example, we find that the presence of light spots in Fig. 11 (top) depends a lot on the camera position; top-left has the least problem. In practice, light spots and reflections can badly influence the performance of eye-related algorithms due to illusions on eyeball and eyelids [1]. Similarly, we find that eyes in the top-left and top cameras are badly occluded by the glass frame (Fig. 11 (middle)). We also observe a vast difference between yawning mouth shapes on all five cameras (Fig. 11 (bottom)). Such observations could be used to design and optimize drowsiness detection algorithms. Better yet, FatigueView contains practical challenges from subtle visual signs and non-drowsiness events (see Fig. 16). Particularly, subtle visual signs are a common occurrence but are barely studied in existing works [7], [62].

In addition to drowsiness-related labels, as presented in Fig. 12, FatigueView also contains distraction-related annotations in the set level. On the one hand, those events bring practical challenges on drowsiness detection since their visual features in the set level are sometimes similar to drowsiness. On the other hand, both drowsiness and distraction annotations are beneficial for control-transferring in self-driving vehicles. This is because resuming control from a highly automated vehicle to manual normally requires attention of a driver on the road ahead [82]. However, drowsiness and engagement in other tasks could badly deprive drivers’ attention to the roadway and may lead to reduced driving performance. For this, timely drowsiness and distraction detection could best inform drivers of their obligation to resume control of driving.

It is worth discussing light conditions in DMS. Theoretically, sunlight intensity, coverage, and angle could potentially influence the performance of drowsiness detection algorithms. For instance, squinting from direct sunlight and RGB image quality reduction in low-illumination environments (e.g. road and tunnel). In practice, images from the IR camera are used for most scenarios since it has robust quality in different light conditions with IR-pods (see Fig. 6). Moreover, drowsiness detection results from the RGB camera are normally ignored in low-illumination environments, e.g. Changan’s UNI-T model. For FatigueView, challenges from different light conditions are properly preserved in both IR and RGB images: (1) our...
Test-Real contains various real drowsiness driving videos with rich sunlight conditions. (2) In Test-Sim, we randomly moved (or covered) the IR pods and overhead lamps to mimic influences from lights (see Section III-C). For research purposes, some existing night-time data augmentation approaches (e.g. Lee et al. [83]) could also be employed to extract various night effects based on RGB images.

D. Limitations

Due to limitations during the global pandemic [84], most participants were bus/taxi/coach/truck drivers and office workers in China. Thus, FatigueView has relatively narrow distribution of anthropological traits. Though reviews from Nazari et al. [85] suggests that there is no evidence showing explicit difference within ethnicity in terms of visual signs and duration, we still believe that it is necessary to evaluate element and set level algorithms in an international setting. For this purpose, existing datasets such as YawDD [48], RLDD [18] and NTHU [1] can be aggregated to represent a more diverse composition of the human population. Moreover, FatigueView will be regularly updated to include participants of various skin, hair and eye colour. Finally, we encourage DMS users and researchers from different countries/cultures to share drowsiness videos (including distraction and real sleepy driving) via FatigueTree in hope that we can continue enriching the diversity of such data amidst our present constraints.

V. BENCHMARK EVALUATIONS

We present a benchmark evaluation on FatigueView with 16 representative approaches, including 3 for eye status classification, 2 for eye blinking recognition, 2 for yawning recognition, 3 for action recognition, and 6 comprehensive methods for drowsiness detection. For fairness, all settings follow their original papers unless stated otherwise. Similar to real conditions in practice, we train and test each approach in a subject- and camera-independent manner. For instance, we only use training data from the RGB-Front camera for model training, and then evaluate it with Test-Sim from the RGB-Front camera. Since Test-Real were mostly recorded in steering column and dashboard, training data from IR-Bottom is employed as their camera position are similar to each other. Table III and IV present the results (Precision $P$ and Recall $R$) on two testing data independently. For comparison, the one producing the highest F1 score ($F = 2PR/(P+R)$) in each group is marked in bold. Summary of advantages and disadvantages of these methods are detailed in Table V.

It should be noted that drowsiness levels are defined by different types and combinations of drowsiness events (see Section IV-B and Fig. 10) in practice. Thus, drowsiness-level detection is actually an event detection/classification task and measured by precision/recall metrics. Overall, we find that values in Table III (IR-Front) and Table IV are slightly different but overall consistent. The results and observations are analysed below.

A. Visual Signs at Set Level

Motivated by the surveys in [6] and [7], we first evaluate seven representative methods on detecting visual signs on face.
However, Dense-LSTM is limited in its ability to recognize yawning from small mouth openings (see Fig. 15 (a) and (b)); we believe this is caused by the model being easily misled by the abundance of speaking frames (negative).

We observe that F1 scores of HOG&LBP in both Table III (IR-Bottom, \( \frac{2 \times 0.70 \times 0.92}{0.70 + 0.92} = 0.8383 \)) and IV (\( \frac{2 \times 0.71 \times 0.90}{0.71 + 0.90} = 0.7938 \)) are obvious lower than its original AOY dataset (0.9277). This is because (1) FatigueView contains various yawning sets representing more practical challenges, and (2) HOG&LBP is not robust enough to deal with complex scenarios found in FatigueView such as occlusion, micro-yawning, etc. Thus, FatigueView provide a better benchmark than existing ones.

**Nodding:** The classic rule-based nodding detection approach HeadMotion [89] is employed as the sole baseline. We find that most failure cases were caused by confusing motions like looking around (or on the phone) and tiny drowsiness-related motions. Since there are differences in the nodding amplitude and direction across the camera groups, using pre-defined rules is not sufficiently robust. As presented in Table III, we also find that both RGB and IR achieved their best performances in the Top-Left camera view, which can be attributed to easy detection of sideway motions. There is no reported result for HeadMotion in Table IV since only one nodding event was observed and annotated in Test-Real (see Table II).

**Stretching:** As no approaches exist for stretching detection, we employ the image sequence-based TSN [91] and skeleton sequence-based 2sAGCN [90] to build baselines as the characteristics of stretching resemble that of action recognition. Both TSN and 2sAGCN performed quite promisingly, especially at specific camera positions (see Table III). However, these methods are not without weaknesses: TSN’s sequence representation is easily impacted by irrelevant features from background and clothing, while 2sAGCN is more robust to obvious stretching (Fig. 15 (c)) but cannot properly recall slight actions like in (d). Such observation is also confirmed in Test-Real: both methods can effectively detect obvious stretching (with recall = 100% in Table IV) while precision of 2sAGCN is 15% higher than TSN.

### B. Drowsiness at Event Level

For drowsiness detection at the event level, three SVM-based methods are first evaluated: Eye4features [20], Eye16Features [17], and FacialUnits [28]. We find that the best results were achieved by these methods, especially when using pre-defined rules.
Table V

<table>
<thead>
<tr>
<th>Methods</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiHiPOG [49]</td>
<td>Combining multiple features with multiple scales</td>
<td>Ad hoc features, limited in generalization</td>
</tr>
<tr>
<td>HiKS [22]</td>
<td>Using multiple features such as PCA and LDA</td>
<td>Ad hoc features, limited in generalization</td>
</tr>
<tr>
<td>EARSequence [86]</td>
<td>Using both EAR and its temporal information</td>
<td>Not robust to head movements such as looking down</td>
</tr>
<tr>
<td>MeanDistance [19]</td>
<td>Optimized EAR, robust to noise, temporal information</td>
<td>Simple classification with limited robustness</td>
</tr>
<tr>
<td>MotionVector [87]</td>
<td>Using optical flow to capture blinking in temporal</td>
<td>Limited in generalization with empirically defined thresholds</td>
</tr>
<tr>
<td>HOG &amp; LBP [33]</td>
<td>Using both HOG and LBP features from mouth and eyes</td>
<td>Low accuracy, especially confusions from normal open mouth</td>
</tr>
<tr>
<td>Dense-LSTM [88]</td>
<td>Fully use the temporal information with LSTM</td>
<td>The first stage is easily overfit to none-drowsiness frames</td>
</tr>
<tr>
<td>HeadMotion [89]</td>
<td>Special tracking using a point between the eyes</td>
<td>Using pre-defined rules is not sufficiently enough</td>
</tr>
<tr>
<td>2xAGCN [90]</td>
<td>Robust action recognition, the model is relatively light</td>
<td>Training data quantity and quality, especially the first one</td>
</tr>
<tr>
<td>TSN [91]</td>
<td>Integrating multiple frames for robust action recognition</td>
<td>Context information from multiple frames are not fully used</td>
</tr>
<tr>
<td>Eye4features [20]</td>
<td>Using blinking frequency and length to detect drowsiness</td>
<td>Facial and body features are ignored, limited in robustness</td>
</tr>
<tr>
<td>Eye16features [17]</td>
<td>Considering more features from eyes</td>
<td>Facial and body features are ignored, limited in robustness</td>
</tr>
<tr>
<td>FacialUnits [28]</td>
<td>Using six drowsiness-related features and their extensions</td>
<td>Not fine-grained enough to capture subtle visual signs</td>
</tr>
<tr>
<td>HM-LSTM [18]</td>
<td>Using four blinking features to learn an LSTM model</td>
<td>Facial and body features are ignored, limited in robustness</td>
</tr>
<tr>
<td>DDDNet [14]</td>
<td>Modeling multiple features from eye, mouth and face</td>
<td>The temporal information is not fully used</td>
</tr>
<tr>
<td>VariousNets [61]</td>
<td>Using features from both multiple frames and optical flow</td>
<td>Context information from multiple frames are not fully used</td>
</tr>
</tbody>
</table>

Eye16features [17], and FacialUnits [28]. Specifically, Eye4features and Eye16features integrate 4 and 16 eye-based features to classify drowsy and normal status, respectively. FacialUnits generates 6 drowsiness-related features based on EAR and MAR (Mouth Aspect Ratio) sequences. The 4 features include blink duration, PERCLOS, blink rate, and percentage of upper and lower eyelid distance changes. The 16 features are mostly extended from the 4, with more statistical values. The 6 features includes PERCLOS, yawning number, blink rate, blink duration, eye closing and reopening speeds. In Table III, we find that Eye16features achieves the best performance, followed by FacialUnits. However, both methods still struggle to properly detect cases of subtle drowsiness in FatigueView since their features are not fine-grained enough to capture subtle visual signs (e.g. Fig. 4 (b)).

For this, we benchmark three methods that are representative of neural networks: Hierarchical Multiscale LSTM (HM-LSTM) [18], DDDNet [14], and VariousNets [61]. HM-LSTM is an end-to-end method which uses blink features to learn an LSTM (Long Short-Term Memory) model. DDDNet consists of two connected networks for face and landmark detection as well as drowsiness classification. VariousNets employs three independent networks (AlexNet [95], VGG-FaceNet [96], FlowImageNet [97]) to learn drowsiness features from video frames and optical-flow images. In both Table III and IV, the HM-LSTM achieves the best result among this family of methods at the event level due to its capability at detecting subtle blinking signs (early stage drowsiness), and this seemed to contribute to overall drowsiness detection. It is obvious that features from face and body (see Fig. 1 and Fig. 15) can be leveraged for better performance.

C. Analysis of Challenges

We provide a concise analysis of the challenges posed by this dataset and some interesting insights for future work.

Challenges from camera positions: For most visual sign and drowsiness event detection approaches, we find that both RGB and IR cameras in the frontal face achieved the overall best performance among all locations. Location-wise accuracies are widely variable– most existing approaches are sensitive to camera locations. It is also interesting to find that the Left and Bottom cameras achieve the second best performance in visual sign sets and drowsiness events, respectively. Naturally, some participants are nodding-off in drowsiness events and it is easier for the bottom camera to capture different facial signs from image sequences. Inspired by this observation, with DMS as an example, we suggest to use the steering column for camera installation in actual practice. Equally necessary are further algorithms to cater to different occlusion, light spots, etc. to improve the robustness of detection.

Challenges from camera types: We further calculate the F1-score of each method in all camera locations and we find that the rankings between IR and RGB cameras are mostly consistent– we surmise that there is a weak interplay between camera types and locations. However, we also find that the global performance from RGB cameras is better than IR. This is because more than half of the participants wore glasses and light spots in IR cameras badly influence the eye-based drowsiness features. Thus, the mitigation of light spots on different glasses should be intensively investigated if IR cameras are employed.

Challenges from various events: As presented in Fig. 16, FatigueView also contains practical challenges from subtle visual signs (top: heavy-laden eyes, squeezing eyelids, touching face and shaking head quickly) and non-drowsiness events (bottom: eating and drinking).
visual signs and non-drowsiness events (e.g., eating, drinking and laughing). Notably, these subtle signs frequently occur in practice but are barely studied [7], [62]. We suggest to incorporate micro-motion-expression recognition and gaze regression approaches to facilitate drowsiness detection. Moreover, additional event recognition could potentially help filter out non-drowsiness events.

VI. CONCLUSION

We present an extensive large-scale multi-camera dataset that is designed to study real-world drowsiness detection during driving scenario. The dataset was collected via a multi-camera platform with novel collection strategies employed to appeal to the challenges of real-world applications. It also contains real sleepy driving data for testing. We provide hierarchical annotations to support various drowsiness-related research tasks, whereby representative baseline methods were benchmarked at various levels. We hope that the release of FatigueView will benefit future research in this area, particularly to address practical real-world concerns. In future work, we will keep extending the dataset in terms of international setting and warning signs associated with human fatigue, such as reacting slowly, feeling impatient and stiff.

ACKNOWLEDGMENT

The authors would like to thank Xiangbin Kong, Hongzhi Zhang, Junkai Sun, Ming Zhao, Wenfeng Wang, Chunshen Feng and Jianwei Niu for their great support in this project.

REFERENCES


Cong Yang received the Ph.D. degree in computer vision and pattern recognition from the University of Siegen, Germany, in 2016. He has been an Associate Professor with Soochow University, China, since 2022. Before that, he was a Post-Doctoral Researcher in the MAGRIT Team with the INRIA, France. Later, he worked scientifically and led the computer vision teams at Clobotics and Horizon Robotics. His main research interests are computer vision, pattern recognition, and their interdisciplinary applications.

Zhenyu Yang received the bachelor’s and master’s degrees in civil engineering from Southeast University, Nanjing, China, in 2016 and 2019, respectively. He is currently a Computer Vision Engineer at Horizon Robotics. His research interests include computer vision, machine learning, and their applications.

Weiyu Li received the bachelor’s degree in software engineering and the master’s degree in computer technology from the Harbin Institute of Technology (HIT), Harbin, China, in 2017 and 2019, respectively. He is currently a Computer Vision Engineer at Horizon Robotics. His current research interests include computer vision and its applications.

John See (Senior Member, IEEE) received the B.Eng., M.Eng.Sc., and Ph.D. degrees from Multimedia University, Malaysia. He is currently an Associate Professor at the School of Mathematical and Computer Sciences, Heriot-Watt University, Malaysia. He is also an Award Recipient of the Belt and Road Initiative Young Scientist Exchange Fellowship, where he was a Visiting Research Fellow with Shanghai Jiao Tong University, China. He has published close to 100 papers in reputable journals and conferences. He is currently serving as the Associate Editor for EURASIP JIVP, IEEE Access, and Frontiers in Signal Processing. His research interests include diverse range of topics in computer vision and pattern recognition, particularly in the emerging fields of facial micro-expressions, affective computing, computational aesthetics, and deep learning.