Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD

Imen Jegham [a], Anouar Ben Khalifa [b,*], Ihsen Alouani [c], Mohamed Ali Mahjoub [b]

[a] Université de Sousse, Institut Supérieur d'Informatique et des Techniques de Communication de H. Sousse, LATIS- Laboratory of Advanced Technology and Intelligent Systems, 4011, Sousse, Tunisia
[b] Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS- Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisia
[c] IEMN-DOAE, Université polytechnique Hauts-de-France, Valenciennes, France

ARTICLE INFO

Keywords:
Safe driving
Intelligent transportation system
Driver distraction
Multiview
Multimodal
Multispectral
Public dataset

ABSTRACT

Driver distraction and fatigue have become one of the leading causes of severe traffic accidents. Hence, driver inattention monitoring systems are crucial. Even with the growing development of advanced driver assistance systems and the introduction of third-level autonomous vehicles, this task is still trending and complex due to challenges such as the illumination change and the dynamic background. To reliably compare and validate driver inattention monitoring methods, a limited number of public datasets are available. In this paper, we put forward a public, well-structured and complete dataset, named Multiview, Multimodal and Multispectral Driver Action Dataset (3MDAD). The dataset is mainly composed of two sets: the first one recorded in daytime and the second one at nighttime. Each set consists of two synchronized data modalities, both from frontal and side views. More than 60 drivers are asked to execute 16 in-vehicle actions under a wide range of naturalistic driving settings. In contrast to other public datasets, 3MDAD presents multiple modalities, spectrums and views under different time and weather conditions. To highlight the utility of our dataset, we independently analyze the driver action recognition results adapted to each modality and those obtained of several combinations of modalities.

1. Introduction

Intelligent Transportation Systems (ITS) are becoming an important component of our society. They are meant to enhance transportation safety, efficiency and sustainability as well as comfortable driving experience [1,2]. One of the main key focus of ITS is the Advanced Driver Assistance System (ADAS) technology, which plays a crucial role in ensuring the vehicle, driver, pedestrian and passenger's safety and comfort. Properly used ADAS technologies can prevent 40% of all vehicle crashes and about 30% of traffic deaths [3]. In the USA, vehicular collisions, in 2016, caused 37,461 fatalities and more than 2.4 million demoralizing injuries, with an estimated cost of 242 billion dollars. Importantly, more than 37,400 people were killed in traffic crashes (a 5% increase from 2015) [4,5]. A high amount of fatalities occurred in darkness or twilight when it was very difficult for drivers to see clearly.

Driver inattention is an extremely influential contributing factor of road crashes and incidents. It is defined as a diminished attention to activities that are critical for safe driving in the absence of a competing activity [6]. This factor can be clustered into two main classes: distraction on the one hand, and fatigue and somnolence on the other hand [7]. This was confirmed by an online survey, in which Peter et al. [4] affirmed that the two main leading contributors to severe crashes were distraction by secondary tasks and poor visibility in low light that generally made drivers sleepy.

With the continuous improvement and development of advanced driver assistance systems up to very automated driving functions, drivers are allowed to engage temporarily in non-driving related tasks. However, they need to react appropriately to a taking control request when the automated vehicle reaches its limitations. To support the driver in such situations, driver inattention monitoring systems might permit adaptive take-over concepts [8]. Thus, monitoring driver inattention is still important and is a trending topic that faces several challenges [9] such as illumination variations, cluttered background, etc. Non driving related tasks are numerous and can take many forms. Thus, the National Highway Traffic Safety Administration categorized distraction into four groups [10]: cognitive distraction when the visual field of the driver is blocked where they have to be looking while driving, visual distraction when the driver neglects looking at areas they should be looking to while driving, physical distraction when both driver's hands (or one hand) are taken off the steering wheel to manipulate an object, and auditory distraction when the driver

[*] Correspondence to: ENISo, 4023, Sousse, Tunisia.
E-mail address: anouar.benkhalifa@eniso.rnu.tn (A. Ben Khalifa).

https://doi.org/10.1016/j.image.2020.115960
Received 4 October 2019; Received in revised form 9 July 2020; Accepted 2 August 2020
Available online 8 August 2020

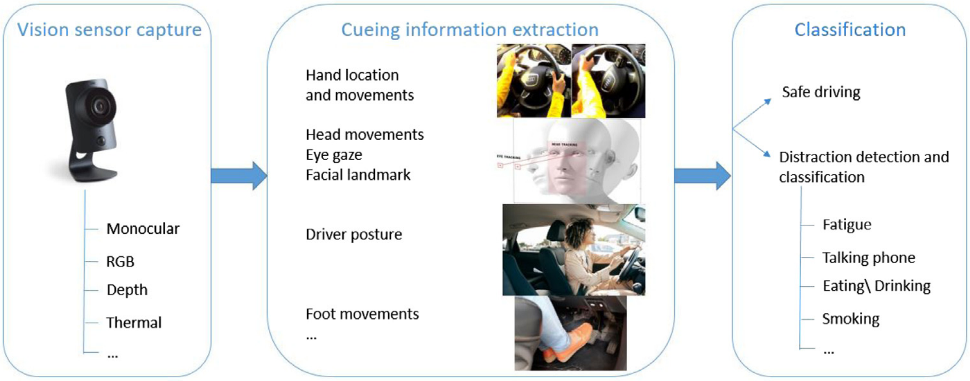

Fig. 1. General overview of basic components of vision-based monitoring driver system.

is prevented by sounds from making the best use of their hearing since their attention is drawn to whatever causes the sound. Most of non driving activities can include more than one of these classes, for example talking to a cell phone which creates cognitive, physical and auditory distraction.

Crash data affirm that poor visibility at night is one of the leading contributors to fatal collisions. In fact, for decades, nighttime fatality rates have been three to four times higher than daytime rates [4]. Multiple factors are involved in the road safety difference between day and night. The main factors are poor visibility coupled with distracted driving and drivers' fatigue that is a natural reaction to darkness. Fatigue does not have a universal definition [11]. In an attempt to avoid accidents, most fatigued and sleepy drivers will try to avoid sleeping. Thus, certain physical and physiological phenomena that precede the onset of sleep can be observed. When a driver is tired and begins to resist getting sleepy, several symptoms can be noticed such as the increased frequency of touching eyes, heads and faces, repeated yawning, the difficulty of keeping eyes open, slower responses and reactions, etc. Consequently, in our study, we consider fatigue and somnolence as a secondary task activity.

To maintain safe driving, monitoring driver inattention is vital and crucial. As a result, several efforts to detect and recognize driver distraction are made utilizing different acquisition devices. To be effective, such sensors need to be human centric and take into account a lot of system components including: driver monitoring (e.g. looking at the driver to recognize their activity and attention state), vehicle sensors (e.g. looking at the vehicle speed, the steering angle, braking, etc.) and vehicle surroundings (e.g. looking at road and other cars to understand the surrounding situation) [12,13]. In this paper, we focus mainly on driver monitoring given that it offers a deeper understanding about recognizing common inattention.

Different physiological and physical signals have been captured such as ECG, EEG, EOG, etc [14–19]. However, their high cost and the complexity of their installation have led to the use of vision sensors. These latter offer the most direct method for detecting the early onset of distraction [7], and there are excellent means of optimization since such sensors can be seen as an excellent platform to be shared with other vision-based driver assistance applications. Multiple types of vision sensors can be used including monocular cameras, depth cameras, etc. Given their clarity, color images prove their robustness, especially in controlled environments. However, in naturalistic driving settings, captured color images are impacted by various weather conditions, which influence the image quality [20,21]. For this reason, researchers tend to supplement or even replace images provided by monochrome and color cameras in the visible spectrum with practical images from other modalities with the intent of improving the performance of the whole system no matter what the weather conditions are like, but still keeping or even ameliorating the types of features and classifiers. At the same time, infrared cameras can produce clear images under the same conditions (day or night), and they are designed to be used in poor lighting and poor weather conditions.

In real-world driving settings, the accuracy of vision-based driver inattention monitoring methods remain limited because of the big number of challenges present in this domain such as dynamic background, occlusion, and bad visibility [22]. To effectively compare these methods, public datasets are crucial. They allow the direct comparison of numerous methods with the state of the art and they open challenging questions to a wider community. Therefore, several datasets have been collected, whether they are in simulated assisted-driving, naturalistic driving settings or a parked vehicle. However, most of them are not publicly available.

In order to facilitate research activities in this field, we propose in this paper a complete, useful and publicly available dataset, named Multiview, Multimodal and Multispectral Driver Action Dataset (3MDAD), which is designed to overcome the limitations of the aforementioned databases. Our new public, multispectral, multimodal and extensive dataset highlights the issues observed in naturalistic driving settings including multiple users, dynamic and cluttered background, varying viewpoints and lighting conditions employing a Kinect camera during both daytime and nighttime. 3MDAD presents an important number of distracted actions reported by the WHO [23]. In daytime, it provides temporally synchronized RGB frames and depth frames. At nighttime, 3MDAD contains temporally synchronized infrared frames and depth frames. Such a dataset is of a valuable benefit to researchers working in different fields like image processing, computer vision, sensors fusion [24,25], and human-centered intelligent driver assistance systems.

The main contributions of this paper include:

(1) Introducing a new extensive, multimodal, multiview and multispectral dataset that highlights the issues observed in naturalistic driving environments, employing two Kinect depth cameras in daytime and at nighttime. The dataset is publicly available.[1]
(2) Exploring public databases to our knowledge while positioning ourselves against these datasets, and providing an overview of some selected private datasets
(3) Assessing the relevance of using two cameras simultaneously in multiple viewpoints where each of them delivers different modality by applying early fusion at both day and night times

The remaining of this paper is organized as follows: In Section 2, the main publicly available datasets and some private datasets related to ours are briefly reviewed. Our new public dataset is described and the main differences with the existing public dataset are pointed out in Section 3. The main naturalistic driving setting challenges are described in Section 4. Section 5 demonstrates the utility of our dataset in recognizing drivers' in-vehicle actions by reporting several experiments

[1] https://sites.google.com/site/benkhalifaanouar1/6-datasets.

Table 1
Overview of some collected private datasets.

Ref	Sensors	S[a]	Actions	Nature of content	Experimental settings	Day or Night	Objective	Monitored body parts
[26]	1 Mono chrome camera	9	3 : Safe driving, Tuning instrument cluster, Tuning the gear shift	9 videos	Naturalistic driving	Day in different weather conditions	Vision-based hand activity analysis to understand driver behavior, in particular as it relates to attentiveness and risk	Hand
[27]	1 Kinect sensor	8	5: Safe driving, Making a call, Drinking, Sending an SMS, Looking at an object inside the vehicle	Set of videos	Simulated environment	In early morning and in late evening	Driver distraction detection and recognition based on eye behavior, arm position, head orientation, and facial expressions	Face
[28]	1 Frontal camera facing driver and road camera facing road	20	7: Tuning radio, GPS operating, GPS following, Dialing phone, Talking phone, Describing picture, and Talking to passenger	Set of videos	Naturalistic driving	On day under good weather condition	Automatic extraction of facial cues from frontal camera facing driver to assess perceived visual and cognitive distraction of drivers performing secondary tasks	Face
[29]	1 Color camera	40	4: Grasping the steering wheel, Operating the shift lever, Talking on a cellar phone and Eating a cake	Images	Parked vehicle	On day in varying illumination	Investigate different pattern classification paradigms to automatically understand and characterize driver behavior	Posture
[30]	2 Cameras	4	3 : Gear region activities, Wheel region activities, Instrument cluster region activities	Set of videos	Naturalistic driving	Day	Driver activity recognition based on head, eye, and hand cues	Head, eye, and hand
[31]	1 Color camera	5	6: Normal driving, Talking, Texting, Eating/Drinking, Hair and Makeup, Reaching	Video frames	Naturalistic driving	On day under varying degrees of lighting	Distracted driving behavior classification and recognition	Posture
[32]	Multiple high resolution cameras and depth sensors observing driver	11	6: Vigilant driving, Gesturing, Talking to a passenger, Operating the infotainment unit, Interacting with a cell-phone or tablet and Drinking a beverage	Set of videos	Naturalistic driving	Day	Driver take-over readiness estimation based on observable cues from in-vehicle vision sensors placed on conditionally autonomous vehicle	Gaze, hand, pose, foot

[a] Subjects.

based on the features extracted from Spatio Temporal Interest Points (STIPs) and automatically extracted features based on deep learning. The conclusion is finally stated in Section 6.

2. Related work

Monitoring driver inattention is one of the most active research areas in both machine learning and computer vision. There are several ways to monitor driver inattention depending on specific purposes. Fig. 1 shows a general overview of the basic components of a common vision-based monitoring driver system. We notice that such systems can use several types of vision sensors, extract different kinds of features that will be used separately or merged, and try to detect and recognize driver distraction. Thus, the work in this field can be grouped into: inattention or distraction detection, and distraction recognition and classification. However, most of these efforts focus mainly on monitoring drivers in the daytime and ignore the nighttime. Given that fatigue is more common at night, researchers have considered fatigue and somnolence as the main distraction. In fact, distracted driving is coupled with limited visibility, which leads to more fatal crashes [4].

Despite the large body of the literature and the big importance of research in monitoring driver inattention, only a limited number of public datasets are available. To validate their proposed approaches, researchers have been committed to follow one of these three alternatives:

- Use of existing public datasets that are very limited
- Use of public datasets that are not dedicated to drivers
- Collection of their own data usually obtained using a parked vehicle or a simulated environment, for security reasons. Even in naturalistic settings, they describe a limited number of real world driving issues. As for our multiple attempts to obtain them, these data are still private. To provide an overview of several selected private datasets, we summarize the most known ones in Table 1 with the main important elements associated with these datasets.

In the following, we review the existing driver public datasets, up to our knowledge.

2.1. Distraction detection

To detect distraction, researchers tend to detect, localize and track some driver body parts such as faces, hands, feet, heads, and gazes. According to the basic rule of safe driving, the driver must keep their hands on the steering wheel and their eyes on the road [9]. For that, hands and heads are considered the key of driver distraction detection. Besides hands and faces, the feet of drivers have a significant role in controlling vehicles. The feet movement before and after a pedal press can provide valuable information for a better semantic understanding of a driver's behavior, state and style. Nevertheless, only feet data are very poor to detect driver distraction and to recognize driver in-vehicle actions. Thus, only datasets of upper driver body parts have been generally recorded.

2.1.1. Driver hand dataset

In the computer vision field, especially in the automotive context, detecting, localizing and tracking driver hands are crucial and have several potential applications including understanding drivers' activities and behaviors, analyzing their alertness, and investigating their attention level. For this reason, hand detection and tracking are the topic of some work and datasets. Borghi *et al.* [33] introduced a dataset called Turms, which consisted of 14k infrared images of driver hands obtained during naturalistic driving activities through a leap motion placed on the back of the steering wheel and oriented to avoid the occlusion of drivers' bodies. However, despite its insensitivity to variations in lighting conditions, this dataset described a very limited location of the hand due to the choice of the leap motion and its position. Das et al. [34] collected an annotated video-based dataset, called the VIVA dataset, for the task of hand detection under challenging naturalistic driving settings. The dataset included images captured using different sources and camera viewpoints. They were gathered mainly from videos recorded from their laboratory and from YouTube videos of drivers who had similar viewpoints as those observed in their testbed imagery. The assembled dataset presented several naturalistic driving issues such as illumination condition variations, non-hand color similarity, and varying viewpoints.

2.1.2. Driver face and head dataset

Driver's face and head monitoring systems have been studied extensively in the vision and learning community. They play an important role in detecting inattentive driving. Driver distraction can be detected from head movements, gaze directions, eye-lid closures, yawning, and blinking. To drive safely, a driver has to keep their eyes on the road. A list of rules are associated with this condition, such as the well-defined position of the head, eyes, etc. To check these latter, a lot of datasets have been recorded. The CVC11 database was recorded and annotated by the CVC ADAS group in 2016 [35]. The dataset contained image sequences of four subjects with several facial features such as glasses and/or a beard while driving in real scenarios. It was composed of 606 samples, acquired over different days. A set of labels assigned each image into three possible gaze direction classes: looking-right, frontal and looking-left. In order to highlight some of the issues in head pose estimation and face detection in a naturalistic vehicle environment in daytime, the VIVA Face dataset was assembled from selected images collected at the LISA laboratory and selected images from naturalistic driving videos from YouTube [36]. The selection was based on facial occlusion with the vehicle sun visor or the driver hands and on harsh lighting conditions. Moreover, to detect drivers' fatigue, Abtahi et al. [37] introduced a Yawning Detection Dataset (YawDD). Since techniques for yawning detection began with finding face localization, the dataset contained videos of 107 driver faces in real and varying illumination conditions. The dataset contained videos of 107 volunteers from various ethnicities, ages and facial characteristics, and in three different situations: normal driving, singing or talking while driving, and yawning while driving. The aforementioned datasets were collected in daytime. However, night driving could be challenging, and, as a result, driver monitoring would be more important. At nighttime, the research community would assume fatigue and drowsiness as the main driver distraction seen that darkness naturally would make drivers sleepy. To overcome shortages, several datasets were recorded. Park et al. [38] propounded the NTH Drowsy Driver Detection dataset which included a wide range of human subjects of varied gender, illumination, poses, and ethnicities in real life fatigue conditions. In simulated driving, the fatigue of 36 drivers was recorded in five different situations: bareface, wearing glasses, wearing sunglasses, night bareface, and night wearing glasses. Each frame of this video dataset was binary labeled: drowsy or nondrowsy. Du et al. [39] put forward a multimodal distraction dataset that included three modalities: car driving information, face expressions and speech. The dataset was recorded in a simulated driving route by 30 subjects. Even with the big importance of research in monitoring driver faces, only few public datasets have been available, which has led some researchers to collect their own private datasets or to use human face databases such as the Caltecth face database [40] and the DROZY database [41].

2.2. Distraction recognition

Studying each body part separately can be a key of driver distraction detection. However, fusing this information and analyzing driver postures can enhance detection and help to recognize distraction [43]. The whole upper body posture is another important cueing information that should be further explored in monitoring driver inattention. Fig. 2 presents some ranges of driver posture movements that can be related

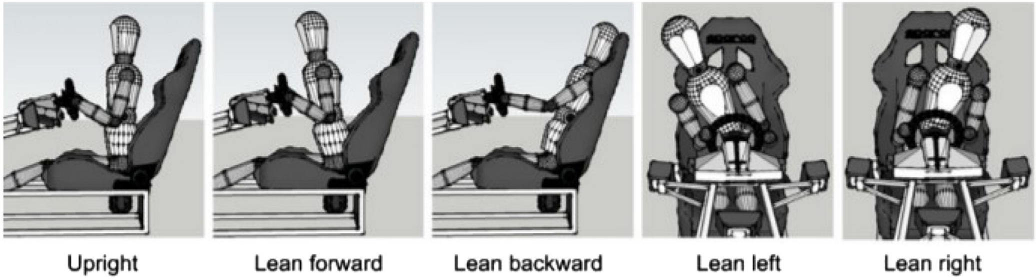

Fig. 2. Illustration of some ranges of driver postures during driving [42].

Table 2
Tasks considered in this study.

Task name	Description
A1: Safe driving	Driver keeps both hands on the steering wheel according to the technique recommended by the National Highway Traffic Safety Administration (NHTSA) known as 9 and 3: considering the wheel as a clock, the left hand on the left portion of the steering wheel in a location approximate to where 9 would be located and the right hand on the right portion of the wheel where 3 would be located.
A2: Doing hair and makeup	Driver looks to interior mirror to do their hair or to put make up.
A3: Adjusting radio	Driver tunes the radio until obtaining predetermined radio station.
A4: GPS operating	Driver inputs the destination address into GPS and then follows GPS instructions.
A5: Writing message using right hand	Driver dials a message using right hand.
A6: Writing message using left hand	Driver dials a message using left hand.
A7: Talking phone using right hand	Driver interacts with the phone call and uses right hand to hold the phone.
A8: Talking phone using left hand	Driver interacts with the phone call and uses left hand to hold the phone.
A9: Having picture	Driver uses cell phone to have picture: taking selfie or pictures of road surroundings.
A10: Talking to passenger	Driver is fully engaged in speech with passengers
A11: Singing or dancing	Driver interacts with music by singing and moving rhythmically.
A12: Fatigue and somnolence	Driver feels tired and exhausted: generally yawning and barely opening eyes.
A13: Drinking using right hand	Driver takes a cup using right hand and drinks a liquid.
A14: Drinking using left hand	Driver takes a cup using left hand and drinks a liquid.
A15: Reaching behind	Driver stretches out their arms in different directions in order to touch or grasp something.
A16: Smoking	Driver smokes a cigarette (for smokers) or imitates this action using an electronic cigarette.

to driver distraction detection; e.g., leaning backwards might indicate a relax position and sleepy driving, and leaning left or right can show a distracted driving, particularly the action: reaching behind.

The level of distraction associated to a given non-driving related task depends on the extent to which a driver is engaged in the task. Different secondary tasks have various concentration requirements. Thus, there are several levels of distraction requiring a varied driving assistance, hence the importance of recognizing and classifying the distracted actions of the driver. However, a restricted number of public datasets are available. The first publicly available "State Farm Dataset" posture classification dataset [44] was launched by the US insurance group to improve its success in a Kaggle contest to automatically detect drivers engaged in secondary activities using dashboard cameras. The dataset performed by twenty-six subjects with different body sizes, ages and gender, in varying illumination. It described ten postures to be classified: safe driving, talking on the phone using right hand, texting using right hand, talking on the phone using left hand, texting using left hand, drinking, talking to passengers, operating the radio, reaching behind, and doing hair and makeup. This dataset was restricted to the purposes of the competition. For this reason, the American university in Cairo, Egypt, was inspired by the State Farm dataset to introduce the AUC distracted driver dataset [45]. This latter was collected in a parked vehicle using an ASUS ZenPhone rear camera fixed using an arm strap to the car roof handle on top of the seat of passengers. The dataset described the same postures as the State Farm dataset of 31 participants from seven countries. These last two databases were composed of images captured arbitrarily. Furthermore, Billah et al. [12] introduced the EBDD video dataset that was composed of video sequences captured by a camera mounted on the front windshield inside a vehicle. The EBDD video dataset was developed by considering the diversity in driving environments as well as the expertise of drivers. Thirteen drivers were called to perform five different activity classes: safe driving, eating, unattentive driving, talking on the phone, and texting.

3. 3MDAD dataset

Motivated by the need for public datasets and the limited number of existing ones, we introduce 3MDAD which addresses multiple aforementioned shortages of the state of the art datasets. This includes the multimodal synchronized data, the diversity of performed in-vehicle actions, and the variety of drivers. The introduced dataset opens challenging questions of real world driving settings to a wider community. Table 2 shows a summary of the existing public datasets compared to 3MDAD.

3.1. Data collection

For the acquisition of our driver dataset, two Microsoft Kinect cameras [46] are used. The first Kinect is placed on the car handle on the top of the passenger window, and the second is mounted on the dashboard in front of the driver. The use of this varied modality sensor is intentional and interesting because of its practicality or relative non-intrusiveness aspect. The Kinect camera consists of a variety of sensors like color (RGB), infrared (IR), and microphones. It is very cheap, widely available, not requiring much computational power, and easy to operate for the real-time manipulation of its generated data.

A good driver assisting system must be effective for day and night vision. Given their clarity in daytime, color images show their efficiency and robustness. However, in real driving settings and under various weather conditions, the RGB image quality can be degraded. Therefore, researchers have tended to supplement or replace images captured by color cameras with practical images from other modalities to improve performances regardless of weather and illumination conditions. The infrared modality, which is designed to be used in poor lighting and poor weather conditions, is one of the most widely used modalities. It can capture clear images under the same day or night conditions. As a consequence, with an acquisition rate of 30 Hz, in daytime, each Kinect

Table 3
Comparison of our introduced 3MDAD datastet to existing public datasets.

Dataset	Sensors	Subj/Act	Nature of content	Experimental settings	Day or Night	Application
Turms [33]	1 Leap motion	7/-	Infrared frames	Naturalistic driving	Day and night	Hand detection and tracking
VIVA Hand Detection [34]	1 Camera in 7 views	-/-	Images	Naturalistic driving	Day	Hand detection
DrivFace database [35]	1 Camera	4/3	Images	Naturalistic driving	Day	Face detection
VIVA Face dataset [36]	1 Camera	-/-	Images	Naturalistic driving	Day	Face detection and head pose estimation
Yawning Detection dataset [37]	1 Camera	107/3	Video sequences	Naturalistic driving	Day	Recognition and tracking of face and mouth
NTH Drowsy Driver Detection [38]	1 Color camera	36/2	Video frames	Simulated settings	Day + night	Face detection
State Farm Dataset [44]	1 Color camera	26/10	Images	Naturalistic driving	Day	Action recognition
AUC distracted driver dataset [45]	1 Color camera	31/10	Images	Parked vehicle	Day	Action recognition
EBDD video dataset [12]	1 Color camera	13/5	Video sequences	Naturalistic driving	Day	Action recognition
3MDAD (our)	2 Kinect cameras	50/16 19/16	Video frames	Naturalistic driving	Day night	Action recognition, face detection, head pose estimation, Hand detection and tracking, passenger detection

Table 4
New 3MDAD dataset summary of images.

	Day	Night
A1	6,761	1,857
A2	6,453	2,046
A3	6,534	1,803
A4	7,657	2,284
A5	6,337	1,916
A6	6,360	1,864
A7	7,373	2,130
A8	7,487	1,987
A9	7,256	2,036
A10	8,400	2,569
A11	7,243	1,951
A12	6,843	2,221
A13	6,387	1,861
A14	6,396	1,843
A15	5,801	1,843
A16	7,738	2,296
Total	444,104[a]	130,028[b]

[a] RGB+Depth+Side view+Front view.

[b] IR+Depth+Side view+Front view.

camera captures an RGB color image with a resolution of 640*480 pixels and a 16-bit depth image. At nighttime, each Kinect camera captures an infrared image with a resolution of 640*480 pixels and a 16-bit depth image. Although every acquired modality couple inside the Kinect is not perfectly synchronized, the temporal difference is not noticeable in the output due to the relatively high frame rate. Fig. 3 demonstrates the layout of used sensors.

3.2. Data description

Our 3MDAD includes real-world driving scenarios during both daytime and nighttime. In daytime, 50 participants (38 males and 12 females) aged between 19 and 41 and at nighttime 19 other participants (11 males and 8 females) aged between 19 and 53 are asked to drive a Volswagen Polo in naturalistic driving settings. Every driver is asked to perform one safe driving and 15 various common secondary tasks under different route segments. Prior to each recording, the participants are given instructions on what action to perform, and no specific details are given on how the action has to be performed. The subjects therefore incorporate different styles in performing the actions. Table 3 describes the selected actions that are commonly executed by individuals while driving. The choice of these activities is proved by the list of the potentially distracting activities delivered by the American Automobile Association Foundation for traffic safety [47] alongside with the study of the database of the fatal accident reporting system collected by the US Department of Transportation [48].

Fig. 4 illustrates the snapshots from all the performed tasks Fig. 4a depicts the dataset tasks in daytime from the side view, while Fig. 4b shows the dataset tasks at nighttime from the front view. All the drivers perform each task for a range of 20 to 35 s, yielding about 1,120 frame sequences (800 in the day and 304 at night) which correspond to about 507 min of the total recording time (367 in the day and 140 at night). Table 4 presents the number of frames for each driver in-vehicle action during the day and the night per view and per modality, where Ai represents the ith action among the ones mentioned in Table 4. The data are collected during the day under varying illumination due to the changing natural conditions and during the night in different weather conditions. A high amount of information can be obtained from multiview, multispectral and multimodal observations (Fig. 5).

4. Dataset design criteria

To be complete, the proposed 3MDAD should contain frame sequences to assess the main issues related to monitoring driver distraction. In this section, we describe the challenges related to naturalistic-driving settings that we strive to represent in our dataset.

4.1. Driver variability

Having a unique model of driver representation is complex and not suitable given the high human variability. This latter comes mainly from the posture, clothing or even the shape of the person, as well as from some external factors such as illumination and partial occlusion. Moreover, due to their anthropometric and anthropomorphic variations, drivers present a huge range of variations in poses and appearances. In fact, depending on the age, gender, body flexibility, looks, angles of view or position in the observed scene, the human being shows variability in sizes, postures and appearances. According to statistics, the bulk of the fatal crash problem will continue to reside with drivers younger than age 65, and particularly with the youngest drivers [49]. The problem lies predominantly in the youngest age groups that have markedly had higher risks for fatal crashes because elderly drivers have low exposure. In addition, each person has a specific body size proportion and owns a comfort zone while executing an action. This latter can appear differently owing to the high flexibility of the human body and the human body skeletal structure that makes different disjointed deformations. In the automotive field, the effects of seat, vehicle package, and anthropometric variables on postures have been largely studied. It has been found that the driver presents a big change in limb postures, while the torso posture remains relatively constant to adapt to changes in vehicle geometry [50]. In our study, we take into consideration all these variations. Our dataset is performed by about 64 drivers of different ages, gender, looks, styles, body sizes, etc as shown in Fig. 6. Prior to each recording, drivers select their seat cushion angle, fore-aft steering wheel position and preferred driving postures in multiple combinations of seat height.

4.2. Illumination variation

Illumination changes such as light variations along the day and sudden illumination changes such as change in weather conditions often cause many problems. Moreover, shadows and highlights cast by moving objects appearing in driving scenes drastically reduce the quality of the results of many driver monitoring algorithms. In naturalistic driving settings, a sudden illumination variation can appear with the background change, its shadow, which is frequent especially with the change in a vehicle direction (Fig. 7).

4.3. In-vehicle action complexity

In naturalistic driving settings, performed in-vehicle driver actions can be extensively categorized into two groups:

- Primary tasks that are important for directing and operating the course of a vehicle
- Non driving related tasks that can be purely distracted such as talking in a cell phone or can be related to the primary task of driving such as moving the head forwards before a lane change to prepare for a better visual check

Hence, even when executed by the same person, the same action appears differently. This is the cause of the high intraclass variability. In addition, seen that the driver action field is restricted, some in-vehicle actions appeared the same from a specific viewpoint. This creates a high interclass similarity.

4.4. Occlusion and camera viewpoint

An occlusion is the temporary disappearance of driver human body parts by being behind another object or passenger of a greater apparent diameter. This issue is common because of the driver look (e.g. wearing sunglasses) as well as the camera position in the vehicle and driver pose. This issue increases when using a single viewpoint. Fig. 8 shows an example of occluded body parts while performing actions. The

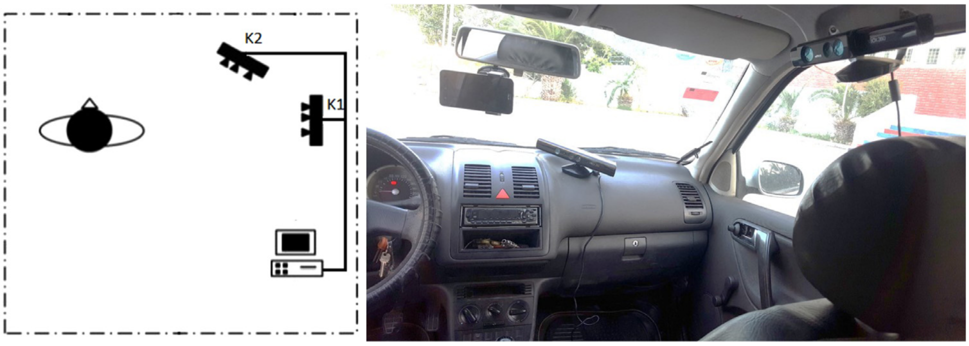

Fig. 3. Layout of data acquisition system.

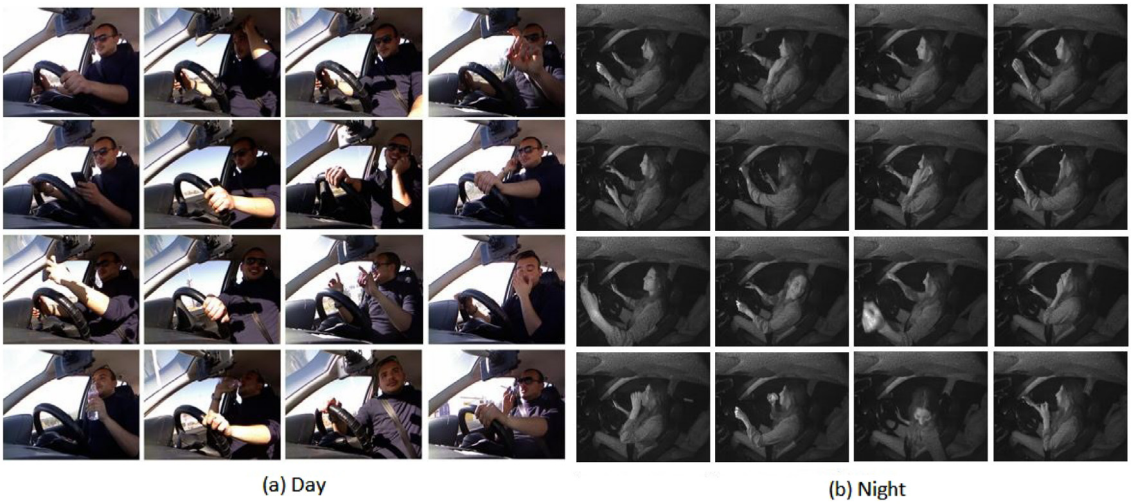

Fig. 4. Snapshots from all tasks available in dataset.

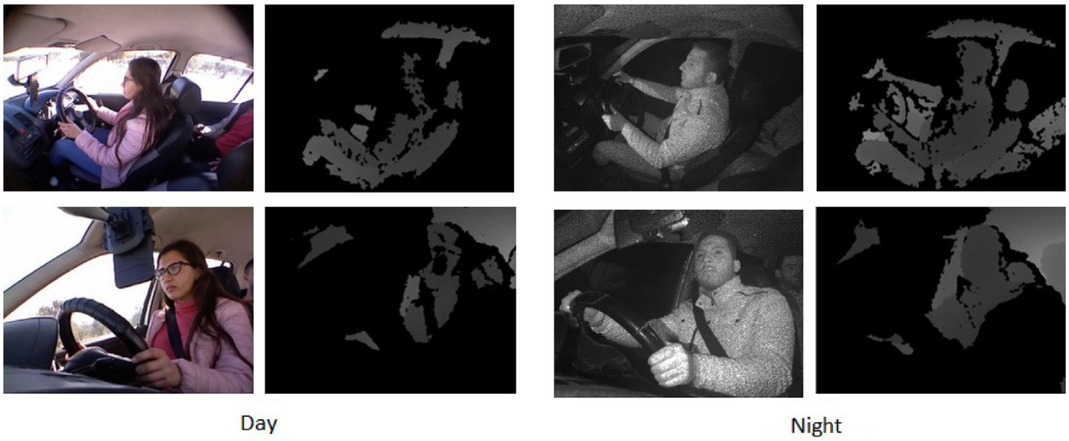

Fig. 5. Samples of each view and each modality.

figures on the left show a tired driver with eyes almost closed and occluded with glasses while the figures on the right present two drivers who write a message using the left hand, which is occluded by the right hand. In addition, the same in-vehicle action can lead to different appearances from various perspectives. Thus, the multiview generates useful data to better analyze drivers' actions. It gives a complete view of the action while removing the occlusion. Although a change in perspective can alter the level of occlusion of body parts, their perceived sizes, and their orientations.

4.5. *Cluttered and dynamic background*

The acquisition environments are very important to reliably monitor the drivers. In naturalistic driving settings, the background is considered as a form of interruption that drastically degrades the driver monitoring task. In fact, in a driver monitoring system, the camera is mounted on a moving vehicle capturing moving background appearing from the car window and static background representing the interior of the vehicle.

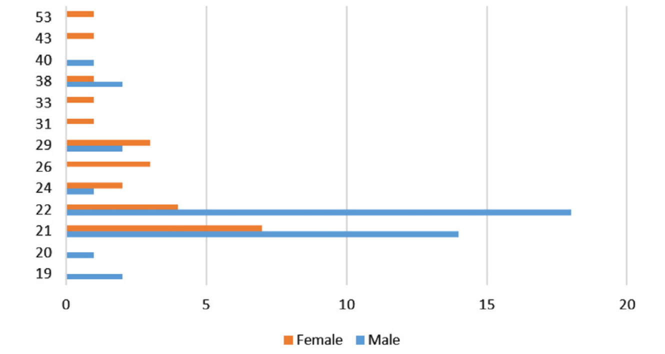

Fig. 6. Distribution of drivers in 3MDAD dataset by age and gender.

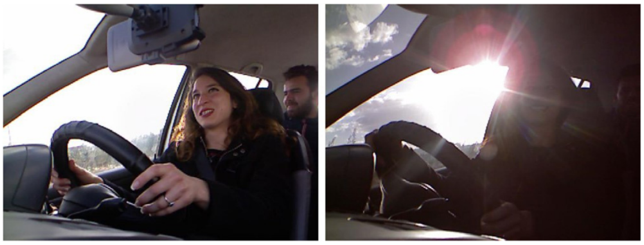

Fig. 7. Illumination variation following change in background.

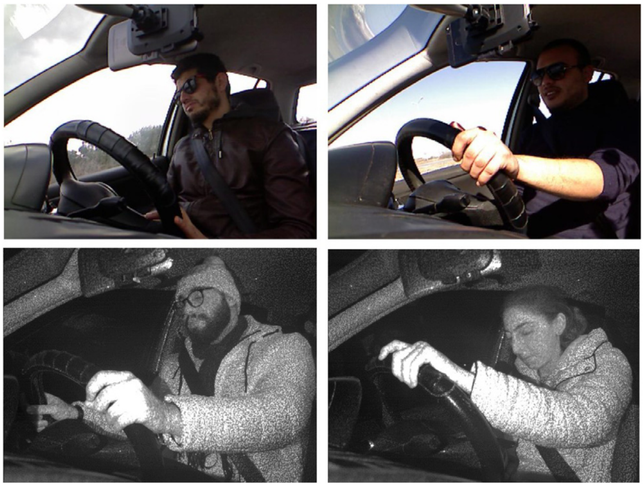

Fig. 8. Occluded body parts while performing actions.

4.5.1. *Static background*

Removing static background in naturalistic driving settings is not an easy task that can be solved by frame subtraction for many reasons, including the gradual change in the appearance of the environment, and the shadows cast by foreground objects. Moreover, background is not completely static but moves irregularly or periodically (e.g. the dynamic background reflected in mirrors). Furthermore, the presence of passengers can dramatically decrease the driver inattention monitoring system performance because of confusion it creates with the driver (Fig. 9).

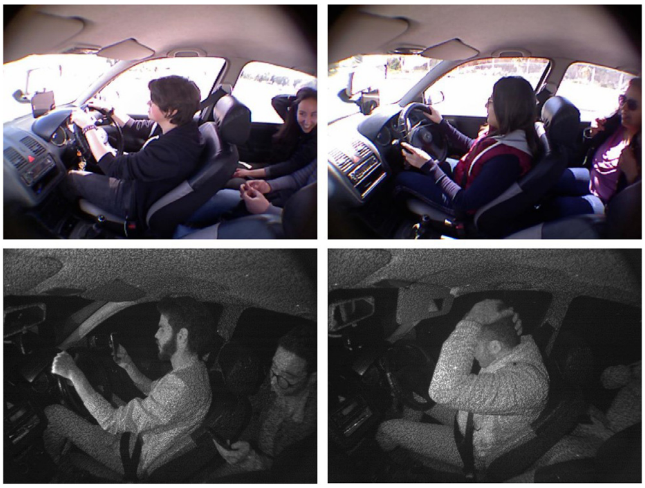

Fig. 9. Moving passengers appearing in camera visual field.

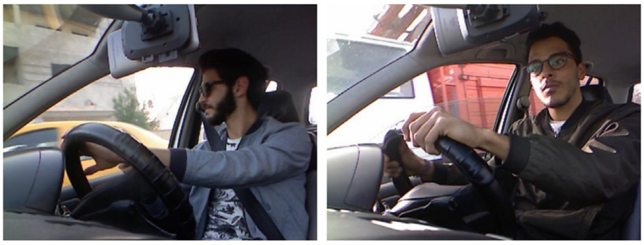

Fig. 10. Cluttered and dynamic background appearing from car windows.

4.5.2. *Dynamic background*

Dynamic scenes are recorded from the window of the car. They can be clustered into two groups: static background including buildings, and traffic lights and dynamic background that move initially at variable speeds including pedestrians, vehicles. In fact, given the position of the camera on a moving vehicle, the background motion is added to the vehicle speed. The modeling of this background remains a complex task (Fig. 10). This issue is partially removed with an infrared modality at nighttime.

5. Driver action recognition experiments

To highlight the utility of our dataset, we choose to perform driver action recognition experiments based on handcrafted features and automatically extracted features.

5.1. *MDAD results*

For the handcrafted features, we employ STIP detection which is the popular video representation and the crucial key step for recognizing actions [51,52]. Spatio-temporal features capture characteristic shapes and motions in videos and provide a compact video representation with respect to their scales and spatio-temporal shifts as well as cluttered background and multiple motions in the scene. In the literature, several methods for STIP feature detection and description have been proposed, and promising action recognition results have been obtained. Wang et al. [53] evaluated and compared suggested space–time features in a common experimental setup across several datasets and proved that the combination of gradient based and optical flow based descriptors extracted from local detected STIPs achieved good results. In the following, we first study the contributions of shape vs motion cues and their fusion for the recognition of actions in each modality and each view. Given multiview and multimodal data, we then opt for feature fusion, and finally we report the action classification results of the most challenging action dataset as well as our proposed one. In all experiments, the k-fold cross-validation technique is employed.

One of the major challenges accompanying the employment of real world videos is the dynamic aspect of the background which appears in our dataset even if our cameras are static. For this reason, after

Table 5
Driver action classification results in terms of recognition rate (%) for each view separately tested for the different modalities.

	Day											
	Side view						Front view					
	RGB			Depth			RGB			Depth		
	HOG	HOF	HOG /HOF	HOG	HOF	HOG /HOF	HOG	HOF	HOG /HOF	HOG	HOF	HOG/HOF
SVM lin	38.02	27.6	37.5	17.18	28.64	29.16	24.47	26.04	34.89	23.95	25	30.72
SVM poly	36.45	25.52	39.06	16.66	23.43	27.08	27.08	20.83	32.29	23.43	19.79	26.04
SVM RBF	33.85	26.04	39.06	23.43	25	30.2	30.02	21.35	34.89	20.83	21.87	29.68
3-NN	18.22	19.27	22.91	13.54	13.54	19.79	15.1	10.93	18.75	11.45	11.97	14.06
5-NN	17.7	18.75	27.08	15.1	19.27	19.79	18.75	14.06	19.27	10.93	14.06	21.35
LRCN	58.85			45.83			48.44			39.58		
	Night											
	Side view						Front view					
	IR			Depth			IR			Depth		
	HOG	HOF	HOG/HOF	HOG	HOF	HOG/HOF	HOG	HOF	HOG/HOF	HOG	HOF	HOG/HOF
SVM lin	37.5	35	45	33.75	28.75	48.75	33.75	31.25	37.5	40	32.5	41.25
SVM poly	37.5	36.25	43.75	35	28.75	46.25	33.75	32.5	36.25	40	32.5	40
SVM RBF	38.75	37.5	47.5	36.25	31.25	45	32.5	30	41.25	37.5	35	40
3-NN	23.75	22.5	23.75	26.25	27.5	25	25	15	22.5	17.5	30	21.25
5-NN	26.25	23.75	28.75	23.75	26.25	27.5	28.75	18.75	28.75	18.75	33.75	27.5
LRCN	72.5			52.5			63.75			46.25		

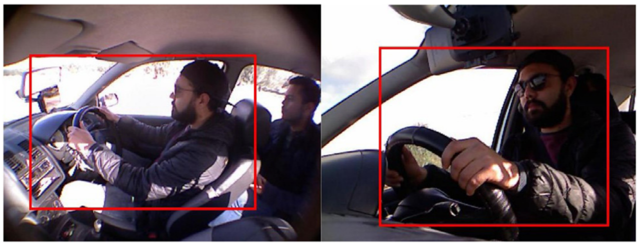

Fig. 11. Extracted regions of interest.

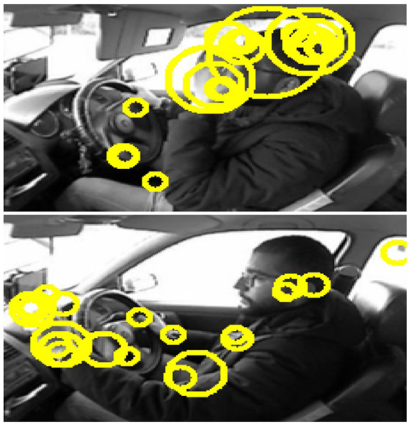

Fig. 12. Space–time interest points detected for two frames with driver actions doing hair and makeup (up) and GPS operating (down).

Table 6
Recognition rates results by integrating multiple fusion using SVM classifier with RBF kernel. For every modality, both side and front views are fused. Accordingly, for every side, we report results of fusing both RGB and Depth modalities.

(a) Day

	Multiview fusion		Multimodal fusion	
	RGB	Depth	Side view	Front view
RR(%)	45.38	41.66	43.75	40.62

(b) Night

	Multiview fusion		Multimodal fusion	
	IR	Depth	Side view	Front view
RR(%)	51.25	52.5	51.25	47.5

observing many captured images, we notice that only the centered window contains moving drivers. This part will be the desired Region Of Interest (ROI). This ROI varies from one view to another. Therefore, we extract the centered 453 × 308 pixels bounding box from frames acquired from the side view and the 463 × 358 pixels bounding box from frames acquired from the front view (Fig. 11).

The role of shape and motion cues for the recognition of biological motions has been the theme of several debates. Computer vision can provide critical insights to this question as various approaches have been proposed. In fact, drivers present different variations in poses and shapes in a way that features extracted from shapes and positions are not sufficient. Thus, researchers have sought to extract motion related features and to combine them with shape features. In this paper, we study and compare the shape HOG descriptor [54] and the motion HOF [55] descriptor separately and the HOG/HOF descriptor of the detected STIPs by the Harris3D detector [56]. Interest points detected for two frames with driver actions are illustrated in Fig. 12.

Given a set of features, we build a spatio-temporal bag of words. This requires the construction of a visual vocabulary. For this reason, we cluster a randomly selected set of STIPs using k-means to build the codebook. The size of our codebook is k=60 which is shown to yield good results after multiple experiments [57]. Finally, we perform the classification experiments using the k-Nearest Neighbors (k-NN) with different values of k as well as the kernel-Support Vector Machine (SVM). For the SVM experiments, the parameters of the kernel are optimized using a search with 25-fold cross-validation. Table 5 summarizes the classification results in terms of recognition rate of various descriptors for each view and each modality for k-NN with k={3,5} and kernel-SVM with RBF, linear and polynomial kernels.

When comparing the obtained results to those obtained in [57], we conclude that the use of ROIs remarkably improves the results due to the decreased impact of cluttered and dynamic background. Moreover, we realize that motion cues alone perform worse than shape cues alone, which is not the case in human action recognition in some previous work [53,58]. This is because of the high complexity of the in-vehicle actions gathered with the high intraclass variability. The combination of descriptors quietly improves the recognition performance.

Given multiview, multimodal, and multispectral data, we choose to fuse features. For multimodal data, we concatenate feature vectors across modalities. Then we concatenate the feature vectors across

Table 7
Summary of some characteristics of different human action datasets and results obtained based on STIP features.

Dataset		Actions	Clips	Back- ground	Resource	Recognition rate
Hollywood2 [59]		12	3,669	Dynamic	Movies, Web, Youtube	32.4%
HMDB51 [58]		51	6,766	Dynamic	Youtube web	20.2%
UCF101 [60]		101	13,320	Dynamic	Youtube	44.5%
3MDAD	Day	16	800	Dynamic	Actor staged	36.97%
	Night		304			44.37%

views. As expected, such fusion ameliorates the recognition rate. Furthermore, the combination of different views yields a complete view of the driver action and reduces or removes occlusion effects. For multimodal fusion, the drawbacks of the features extracted from one modality are revised by features extracted from the other one. Table 6 summarizes driver action recognition results for several combinations where RR denotes the recognition rate. Given that the SVM classifier with the RBF kernel gives generally the best results according to Table 5, we use it for all the multimodal experiments.

Due to their proven efficiency, deep learning techniques are deployed in a wide range of complex real-life problems. Particularly, action recognition applications achieved encouraging progress based on Convolutional Neural Networks (CNNs) and Long Short Term Memories (LSTMs). For this reason, we conduct experiments based on Long-term Recurrent Convolutional Networks (LRCNs) on RGB frames [61] to exploit end to end training of the two mentioned networks. LRCNs combine a deep hierarchical feature extractor (CNN) with a model that can synthesize temporal dynamics. To extract features, we use a pre-trained VGG16 model [62]. Features obtained from the convolution layers consist of 4096 features from each frame. To classify these features, LSTM is able to remember the information of previous inputs and consequently model the temporal sequence. Each LSTM block has memory of antecedent network status, updates hidden states and forgets previous hidden states [63]. The results of the proposed combination of CNN and LSTM are exposed in Table 5. These obtained results show a higher performance than handcrafted features even when using multimodal and multiview fusion. Thus, we believe that this dataset allows the enhancement of deep learning-based action recognition and particularly using multimodal, multiview and/or multispectral fusion.

5.2. Comparison with other human action recognition datasets

Since there is a lack of public driver action dataset in real-world settings, we compare 3MDAD dataset with existing real-world and most challenging human action datasets. The selected datasets present a single modality (RGB) in daytime acquired from multiple views. Therefore, we consider only the RGB modality in daytime and the infrared modality at nighttime from our dataset. Table 7 presents some characteristics of each dataset along with the obtained results using one of the most widely used technique in action recognition experiments based on handcrafted features (STIP features). Low recognition rates are recorded which can be explained by the high amount of challenges present in these datasets including illumination variations and cluttered and dynamic background.

6. Conclusion

The increase in the vehicle infotainment prevalence and busy life drastically decreases driver attention, which makes distraction one of the leading causes of severe vehicle crashes, hence causing death. To validate driver monitoring inattention approaches, a new public, complete, well structured dataset is introduced. The novel dataset contains two sets of naturalistic driving frame sequences, during daytime and nighttime. It is composed of a high variety of frame sequences recorded on different trip routes in Sousse, Tunisia. Each sequence addresses many issues such as dynamic background, action complexity, illumination variations, etc. In contrast to other public datasets, each sequence is annotated by a subject and an action, which makes our 3MDAD not only useful for action recognition, head and hand tracking and localization but also suitable for person recognition tasks. To highlight the utility of our dataset, we have performed action recognition experiments using classical and well known action recognition methods based on STIP features as well as a deep learning technique. The obtained results highlight the high amount of challenges present when compared to the results obtained using the most challenging human action recognition datasets. Moreover, the combination of features from multiple modalities affords better recognition performances.

CRediT authorship contribution statement

Imen Jegham: Conceptualization, Data curation, Writing - original draft, Software. **Anouar Ben Khalifa:** Methodology Visualization, Supervision, Investigation, Software, Validation. **Ihsen Alouani:** Conceptualization, Methodology, Writing - review & editing Validation. **Mohamed Ali Mahjoub:** Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the drivers who have agreed to participate in the collection of the dataset for their support and encouragement.

References

[1] R.P. Loce, E.A. Bernal, W. Wu, R. Bala, Computer vision in roadway transportation systems: a survey, J. Electron. Imaging 22 (4) (2013) 1–24, http://dx.doi.org/10.1117/1.JEI.22.4.041121.

[2] A.B. Khalifa, I. Alouani, M.A. Mahjoub, N.E.B. Amara, Pedestrian detection using a moving camera: A novel framework for foreground detection, Cogn. Syst. Res. 60 (2020) 77–96, http://dx.doi.org/10.1016/j.cogsys.2019.12.003, URL http://www.sciencedirect.com/science/article/pii/S1389041719305212.

[3] A. McDonald, C. Carney, D.V. McGeheet, Vehicle owners' experiences with and reactions to advanced driver assistance systems, 2018, https://aaafoundation.org.

[4] P. Mikoski, G. Zlupko, D.A. Owens, Drivers' assessments of the risks of distraction, poor visibility at night, and safety-related behaviors of themselves and other drivers, Trans. Res. F 62 (2019) 416–434, http://dx.doi.org/10.1016/j.trf.2019.01.011, URL http://www.sciencedirect.com/science/article/pii/S1369847818306508.

[5] A. Mimouna, I. Alouani, A. Ben Khalifa, Y. El Hillali, A. Taleb-Ahmed, A. Menhaj, A. Ouahabi, N. Essoukri Ben Amara, OLIMP: A heterogeneous multimodal dataset for advanced environment perception, Electronics 9 (4) (2020) 560, http://dx.doi.org/10.3390/electronics9040560, URL https://www.mdpi.com/2079-9292/9/4/560.

[6] J.D. Lee, K.L. Young, M.A. Regan, Defining driver distraction, Driver Distract. Theory Effects Mitigat. 13 (4) (2008) 31–40.

[7] Y. Dong, Z. Hu, K. Uchimura, N. Murayama, Driver inattention monitoring system for intelligent vehicles: A review, in: 2009 IEEE Intelligent Vehicles Symposium, 2009, pp. 875–880, http://dx.doi.org/10.1109/IVS.2009.5164395.

[8] T. Pech, S. Enhuber, B. Wandtner, G. Schmidt, G. Wanielik, Real time recognition of non-driving related tasks in the context of highly automated driving, in: International Forum on Advanced Microsystems for Automotive Applications, Springer, 2018, pp. 43–55.

[9] I. Jegham, A. Ben Khalifa, I. Alouani, M.A. Mahjoub, Safe driving : Driver action recognition using SURF keypoints, in: 2018 30th International Conference on Microelectronics (ICM), 2018, pp. 60–63, http://dx.doi.org/10.1109/ICM.2018.8704009.

[10] T.A. Ranney, W.R. Garrott, M.J. Goodman, NHTSA Driver Distraction Research: Past, Present, and Future, Technical Report, SAE Technical Paper, 2001.

[11] M. Cardoso, F. Fulton, J.P. Callaghan, M. Johnson, W.J. Albert, A pre/post evaluation of fatigue, stress and vigilance amongst commercially licensed truck drivers performing a prolonged driving task, Int. J. Occup. Safety Ergon. (2018) 1–11.

[12] T. Billah, S.M.M. Rahman, M.O. Ahmad, M.N.S. Swamy, Recognizing distractions for assistive driving by tracking body parts, IEEE Trans. Circuits Syst. Video Technol. 29 (4) (2019) 1048–1062, http://dx.doi.org/10.1109/TCSVT.2018.2818407.

[13] C. Tran, M.M. Trivedi, Driver assistance for "keeping hands on the wheel and eyes on the road", in: 2009 IEEE International Conference on Vehicular Electronics and Safety (ICVES), 2009, pp. 97–101, http://dx.doi.org/10.1109/ICVES.2009.5400235.

[14] I. Belakhdar, W. Kaaniche, R. Djemal, B. Ouni, Single-channel-based automatic drowsiness detection architecture with a reduced number of EEG features, Microprocess. Microsyst. 58 (2018) 13–23, http://dx.doi.org/10.1016/j.micpro.2018.02.004, URL http://www.sciencedirect.com/science/article/pii/S0141933117303903.

[15] A. Reyes-Muñoz, M. Domingo, M. López-Trinidad, J. Delgado, Integration of body sensor networks and vehicular ad-hoc networks for traffic safety, Sensors 16 (1) (2016) 107.

[16] S. Ameur, A.B. Khalifa, M.S. Bouhlel, A comprehensive leap motion database for hand gesture recognition, in: 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2016, pp. 514–519, http://dx.doi.org/10.1109/SETIT.2016.7939924.

[17] S. Ameur, A.B. Khalifa, M.S. Bouhlel, A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion, Entertain. Comput. 35 (2020) 100373, http://dx.doi.org/10.1016/j.entcom.2020.100373, URL http://www.sciencedirect.com/science/article/pii/S1875952120300811.

[18] S. Jafarnejad, G. Castignani, T. Engel, Non-intrusive distracted driving detection based on driving sensing data, in: VEHITS, 2018, pp. 178–186.

[19] A. Mimouna, A.B. Khalifa, N.E. Ben Amara, Human action recognition using triaxial accelerometer data: Selective approach, in: 2018 15th International Multi-Conference on Systems, Signals Devices (SSD), 2018, pp. 491–496, http://dx.doi.org/10.1109/SSD.2018.8570429.

[20] I. Jegham, A. Ben Khalifa, Pedestrian detection in poor weather conditions using moving camera, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), 2017, pp. 358–362, http://dx.doi.org/10.1109/AICCSA.2017.35.

[21] K. Chebli, A.B. Khalifa, Pedestrian detection based on background compensation with block-matching algorithm, in: 2018 15th International Multi-Conference on Systems, Signals Devices (SSD), 2018, pp. 497–501, http://dx.doi.org/10.1109/SSD.2018.8570499.

[22] I. Jegham, A.B. Khalifa, I. Alouani, M.A. Mahjoub, Vision-based human action recognition: An overview and real world challenges, Forensic Sci. Int.: Digital Invest. 32 (2020) 200901, http://dx.doi.org/10.1016/j.fsidi.2019.200901, URL http://www.sciencedirect.com/science/article/pii/S174228761930283X.

[23] WHO, Distracted driving, 2018, https://www.who.int/violence-injury-prevention/publications/road-traffic/distracted-driving-en.pdf.

[24] W. Lejmi, A. Ben Khalifa, M.A. Mahjoub, Fusion strategies for recognition of violence actions, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), 2017, pp. 178–183, http://dx.doi.org/10.1109/AICCSA.2017.193.

[25] Z. Gao, H. Xuan, H. Zhang, S. Wan, K.R. Choo, Adaptive fusion and category-level dictionary learning model for multiview human action recognition, IEEE Internet Things J. 6 (6) (2019) 9280–9293, http://dx.doi.org/10.1109/JIOT.2019.2911669.

[26] E. Ohn-Bar, S. Martin, M. Trivedi, Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies, J. Electron. Imaging 22 (4) (2013) 1–11, http://dx.doi.org/10.1117/1.JEI.22.4.041119.

[27] C. Craye, F. Karray, Driver distraction detection and recognition using RGB-d sensor, 2015, arXiv preprint arXiv:1502.00250.

[28] N. Li, C. Busso, Analysis of facial features of drivers under cognitive and visual distractions, in: 2013 IEEE International Conference on Multimedia and Expo (ICME), 2013, pp. 1–6, http://dx.doi.org/10.1109/ICME.2013.6607575.

[29] C. Zhao, B. Zhang, J. Lian, J. He, T. Lin, X. Zhang, Classification of driving postures by support vector machines, in: 2011 Sixth International Conference on Image and Graphics, 2011, pp. 926–930, http://dx.doi.org/10.1109/ICIG.2011.184.

[30] E. Ohn-Bar, S. Martin, A. Tawari, M.M. Trivedi, Head, eye, and hand patterns for driver activity recognition, in: 2014 22nd International Conference on Pattern Recognition, 2014, pp. 660–665, http://dx.doi.org/10.1109/ICPR.2014.124.

[31] C. Streiffer, R. Raghavendra, T. Benson, M. Srivatsa, Darnet: a deep learning solution for distracted driving detection, in: Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial Track, ACM, 2017, pp. 22–28.

[32] N. Deo, M.M. Trivedi, Looking at the driver/rider in autonomous vehicles to predict take-over readiness, 2018, arXiv preprint arXiv:1811.06047.

[33] G. Borghi, E. Frigieri, R. Vezzani, R. Cucchiara, Hands on the wheel: A dataset for driver hand detection and tracking, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 564–570, http://dx.doi.org/10.1109/FG.2018.00090.

[34] N. Das, E. Ohn-Bar, M.M. Trivedi, On performance evaluation of driver hand detection algorithms: challenges, dataset, and metrics, in: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015, pp. 2953–2958, http://dx.doi.org/10.1109/ITSC.2015.473.

[35] StateFarm, CVC11: Driver Face dataset (DrivFacce), 2016, URL https://www.kaggle.com/c/state-farm-distracted-driver-detection/.

[36] S. Martin, K. Yuen, M.M. Trivedi, Vision for intelligent vehicles & applications (viva): Face detection and head pose challenge, in: 2016 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2016, pp. 1010–1014.

[37] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, B. Hariri, awDD: A yawning detection dataset, in: Proceedings of the 5th ACM Multimedia Systems Conference, ACM, 2014, pp. 24–28.

[38] S. Park, F. Pan, S. Kang, C.D. Yoo, Driver drowsiness detection system based on feature representation learning using various deep networks, in: Asian Conference on Computer Vision, Springer, 2016, pp. 154–164.

[39] Y. Du, C. Raman, A.W. Black, L.-P. Morency, M. Eskenazi, Multimodal polynomial fusion for detecting driver distraction, 2018, arXiv preprint arXiv:1810.10565.

[40] D. Gong, K. Kwak, Face detection and status analysis algorithms in day and night enivironments, in: 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017, pp. 1–4, http://dx.doi.org/10.1109/ICAICTA.2017.8090965.

[41] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, J. Jang, Real-time driver drowsiness detection for embedded system using model compression of deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 121–128.

[42] C. Tran, M.M. Trivedi, Vision for driver assistance: Looking at people in a vehicle, in: Visual Analysis of Humans, Springer, 2011, pp. 597–614.

[43] Z. Gao, H. Zhang, G. Xu, Y. Xue, A. Hauptmann, Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition, Signal Process. 112 (2015) 83–97, http://dx.doi.org/10.1016/j.sigpro.2014.08.034, URL http://www.sciencedirect.com/science/article/pii/S0165168414003983.

[44] StateFarm, State farm distracted driver detection, 2016, URL https://www.kaggle.com/c/state-farm-distracted-driver-detection/.

[45] H.M. Eraqi, Y. Abouelnaga, M.H. Saad, M.N. Moustafa, Driver distraction identification with an ensemble of convolutional neural networks, J. Adv. Transp. 2019 (2019) 1–12, http://dx.doi.org/10.1155/2019/4125865, URL https://www.hindawi.com/journals/jat/2019/4125865/.

[46] Microsoft, Kinect for windows, 2016, URL http://www.microsoft.com/en-us/kinectforwindows/.

[47] J.C. Stutts, D.W. Reinfurt, L. Staplin, E. Rodgman, et al., The Role of Driver Distraction in Traffic Crashes, AAA Foundation for Traffic Safety, 2001.

[48] L. Qin, Z.R. Li, Z. Chen, M.A. Bill, D.A. Noyce, Understanding driver distractions in fatal crashes: An exploratory empirical analysis, J. Saf. Res. 69 (2019) 23–31, http://dx.doi.org/10.1016/j.jsr.2019.01.004, URL http://www.sciencedirect.com/science/article/pii/S0022437518300574.

[49] L.-H. Chen, S.P. Baker, E.R. Braver, G. Li, Carrying passengers as a risk factor for crashes fatal to 16-and 17-year-old drivers, JAMA 283 (12) (2000) 1578–1582.

[50] M.P. Reed, M.A. Manary, C.A. Flannagan, L.W. Schneider, Effects of vehicle interior geometry and anthropometric variables on automobile driving posture, Human Factors 42 (4) (2000) 541–552.

[51] D. Das Dawn, S.H. Shaikh, A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector, Vis. Comput. 32 (3) (2016) 289–306, http://dx.doi.org/10.1007/s00371-015-1066-2.

[52] B. Lin, B. Fang, W. Yang, J. Qian, Human action recognition based on spatio-temporal three-dimensional scattering transform descriptor and an improved VLAD feature encoding algorithm, Neurocomputing 348 (2019) 145–157, http://dx.doi.org/10.1016/j.neucom.2018.05.121, URL http://www.sciencedirect.com/science/article/pii/S0925231218312876.

[53] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: A. Cavallaro, S. Prince, D. Alexander (Eds.), BMVC 2009 - British Machine Vision Conference, BMVA Press, London, United Kingdom, 2009, pp. 124.1–124.11, http://dx.doi.org/10.5244/C.23.124, URL https://hal.inria.fr/inria-00439769.

[54] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 2005, pp. 886–893, http://dx.doi.org/10.1109/CVPR.2005.177.

[55] S. Danafar, N. Gheissari, Action recognition for surveillance applications using optic flow and SVM, in: Asian Conference on Computer Vision, Springer, 2007, pp. 457–466.

[56] Laptev, Lindeberg, Space-time interest points, in: Proceedings Ninth IEEE International Conference on Computer Vision, 2003, pp. 432–439 vol.1, http://dx.doi.org/10.1109/ICCV.2003.1238378.

[57] I. Jegham, A. Ben Khalifa, I. Alouani, M.A. Mahjoub, MDAD: A multimodal and multiview in-vehicle driver action dataset, in: Computer Analysis of Images and Patterns, Springer International Publishing, Cham, 2019, pp. 518–529.

[58] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: 2011 International Conference on Computer Vision, 2011, pp. 2556–2563, http://dx.doi.org/10.1109/ICCV.2011.6126543.

[59] M. Marszałek, I. Laptev, C. Schmid, Actions in context, in: IEEE Conference on Computer Vision & Pattern Recognition, 2009.

[60] K. Soomro, A.R. Zamir, M. Shah, UCF101: A Dataset of 101 human actions classes from videos in the wild, 2012, CoRR. URL arXiv:1212.0402.

[61] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[62] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[63] A. Graves, Supervised sequence labelling, in: Supervised Sequence Labelling with Recurrent Neural Networks, Springer, 2012, pp. 5–13.