IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 26, 2019, accepted May 12, 2019, date of publication May 16, 2019, date of current version May 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917213

Video-Based Abnormal Driving Behavior Detection via Deep Learning Fusions

WEI HUANG[1], XI LIU[1], MINGYUAN LUO[1], PENG ZHANG[2], WEI WANG[3], AND JIN WANG[4,5]

[1]School of Information Engineering, Nanchang University, Nanchang 330031, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an 710065, China
[3]School of Economics and Management, Chang'an University, Xi'an 710064, China
[4]School of Information Engineering, Yangzhou University, Yangzhou 225000, China
[5]Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, School of Computer & Communication Engineering, Changsha University of Science and Technology, Changsha 410008, China

Corresponding author: Jin Wang (jinwang@yzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61862043, and in part by the Natural Science Foundation of Jiangxi Province under Grant 20181ACB20006 and Grant 20171ACB21017.

ABSTRACT Video-based abnormal driving behavior detection is becoming more and more popular for the time being, as it is highly important in ensuring safeties of drivers and passengers in the vehicle, and it is an essential step in realizing automatic driving at the current stage. Thanks to recent developments in deep learning techniques, this challenging detection task can be largely facilitated via the prominent generalization capability of sophisticated deep learning models as well as large volumes of video clips which are indispensable for thoroughly training these data-driven deep learning models. In this paper, deep learning fusion techniques are emphasized, and three novel deep learning-based fusion models inspired by the recently proposed and popular densely connected convolutional network (DenseNet) are introduced, to fulfill the video-based abnormal driving behavior detection task for the first time. These three new deep learning-based fusion models are named as the wide group densely (WGD) network, the wide group residual densely (WGRD) network, and the alternative wide group residual densely (AWGRD) network, respectively. Technically, WGD takes important issues of deep learning models, i.e., the depth, the width and the cardinality, into consideration when designing its model structure based on DenseNet. For the WGRD and AWGRD, they are more sophisticated as the important idea of residual networks with superpositions of previous layers is incorporated. The extensive experiments are conducted to verify the effectiveness of three new models. Their superiority has been suggested based on rigorous comparisons towards several popular deep learning models in this video-based abnormal driving behavior detection study.

INDEX TERMS Artificial intelligence, digital images, vehicle driving, abnormal driving detection, densely connected convolutional networks, deep learning.

I. INTRODUCTION

It is widely acknowledged that, high-resolution videos are more and more commonly seen within a great number of visual applications at the current stage. For instance, in video surveillance, multiple high-resolution cameras are necessary to be placed at different locations. They work together to identify [1], [2], re-identity [3], [4], and track the moving target [5], [6], making the later high-level analyses based on the moving target (e.g., behavior or even potential intention) more feasible. In emotional computation, high-resolution cameras need to be utilized to capture both obvious and fine changes of emotions of the target person in real-time [7], [8], which has significant impacts in security issues nowadays. It is easy to perceive from the above descriptions that, acquiring and storing a large volume of high-resolution videos are often not difficult to be realized for the time being. However, the main challenge resides in how to efficiently and effectively make correct high-level decisions based on those low-level video clips of large volumes. In this study, high-resolution videos of drivers recorded within vehicles are emphasized. The high-level decision here is to correctly detect abnormal driving behavior (i.e., patterns) of drivers.

Automatic abnormal driving behavior detection is generally accepted as the first issue in realizing the popular fully autonomous driving task. It is certain that, for the autonomous

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

driving task, safety issues are undoubtedly first priorities. It is widely known that, behavior of drivers need to be well restricted in order to avoid any potential accident. Therefore, multiple high-resolution cameras equipped within the driver's vehicle can be utilized to monitor the driver's status in real time. Generally speaking, videos captured by high-resolution cameras also need to be processed immediately, in order to determine whether the current status of the driver is normal or not. It can be acknowledged from the above descriptions that, both the effectiveness (i.e., the detection accuracy) and the efficiency (i.e., the detection speed) of abnormal driving behavior detection are highly demanded. Also, high-speed wireless transmissions are necessary to realize the swift and reliable transmission of high-quality videos, which further facilitates the above automatic abnormal driving behavior detection task [9]– [23].

In order to detect abnormal behavior of drivers, an official and precise definition of abnormal driving is often necessary. According to the International Organization for Standardization (ISO), abnormal driving is defined as the phenomenon that a driver's ability to drive is impaired due to her / his own focus on other activities unrelated to normal driving. Generally speaking, abnormal driving behavior can be mainly divided into three categories. The first one belongs to distracting driving behavior that meets the driver's physical comfort requirements, including smoking, drinking, eating, configuring the aircon, etc. The second one is to meet the driver's need for distracting driving behavior, including makeup, shaving, chatting, using mobile phones or other unnecessary devices, etc. The third one contains distracted driving behavior caused by the surrounding environment, including caring for children, long-term attentions to unexpected events outside the vehicle, etc. Among the above-mentioned abnormal driving behavior, it is necessary to highlight that, the use of mobile phones has already become a major factor in contemporary abnormal driving. In a recent simulation, researchers have found that making a call while driving can cause the driver to distract 20% of her / his attention. More seriously, if the content of the call is important, it will lead to a distraction up to 37%, which will make the driver 23 times more likely to have an accident than normal drivers [24]. Therefore, the use of mobile phones is also considered as one important abnormal driving behavior for automatic detection in this study.

In this study, a single visible-light camera is utilized to record high-resolution videos of the driver, and three novel deep learning-based fusion models are proposed to fulfill the video-based abnormal driving behavior detection task. The basic architecture of deep learning models introduced in this study is mainly motivated by densely connected convolutional networks (DenseNet), which were proposed in 2017 and won the best paper of CVPR the same year [25]. Generally speaking, DenseNet can be regarded as a relatively new convolutional neural network (CNN)-based deep learning architecture, and it has significant merits of reaching the state-of-the-art performance in several well-known classification challenges (e.g., CIFAR, SVHN, ImageNet databases) using less parameters. Also, it is not difficult to be trained even within a tremendously deep model's structure because of its intensive utilization of the residual network [26]. Additionally, deep learning fusion techniques are utilized in this study based on the original DenseNet, in order to obtain three novel fusion models for realizing the video-based abnormal driving behavior detection task for the first time. The three new fusion models proposed in this study are named as the wide group densely (WGD) network, the wide group residual densely (WGRD) network, and the alternative wide group residual densely (AWGRD) network, respectively. Technically, WGD takes important issues of deep learning models, i.e., the depth, the width and the cardinality, into consideration when designing its model structure based on DenseNet. For WGRD and AWGRD, they are more sophisticated as the important idea of residual networks with superpositions of previous layers is incorporated.

The organization of this paper is as follows. In Section II, related works in abnormal driving detection as well as recently popular deep learning studies are briefly reviewed. In Section III, technical details of three novel deep learning-based fusion models are elaborated. In Section IV, extensive experiments based on a large abnormal driving database are conducted and comprehensive analyses are applied. The superiority of newly proposed models are substantiated via rigorous comparisons towards several popular deep learning models implemented in the same task, from the statistical point of view. In Section V, the conclusion of this study is drawn.

Main contributions of this study can be summarized as follows. First, it is the first attempt to incorporate the recently proposed DenseNet into the challenging video-based abnormal driving behavior detection task. Second, technical novelties within newly introduced models of this study, including the important enhancement of width and cardinality in WGD, the sophisticated integration of ResNet and DenseNet in WGRD and AWGRD, are significant. Third, extensive experiments and comprehensive analyses further substantiate the superiority of newly introduced models in tackling the abnormal driving behavior detection problem of this study.

II. RELATED WORKS

In the following, abnormal driving detection and deep learning techniques, which are closely related to this study, are emphasized. Recent developments in the two aspects are briefly reviewed, with pros and cons been discussed.

A. ABNORMAL DRIVING DETECTION

It can be summarized based on literatures of automatic abnormal driving behavior detection that, there are often three commonly used detection schemes. The first one is based on the detection of human physiological signals (i.e., electro-oculogram, electro-encephalogram, respiratory, blood flow changes, etc.) using diverse kinds of sensors [27], [28]. The second one is based on facial details [29] (i.e., changes

in eye movement, mouth movement, head movement, hand features, etc.). The third one is based on motion characteristics of the steering wheel, which is capable to detect the driver's hand pressure [30], the steering time, the brake behavior [31], etc. It is also necessary to point out that, detecting human physiological signals has good real-time performance and high precision, but its main advantage of affecting drivers' normal drivings cannot be neglected, either. Furthermore, physiological signals of human beings vary greatly due to the physiological difference in each individual person and her / his environmental conditions. Therefore, it is challenging to provide quantitative and objective standards for detecting human beings' physiological signals as well. For detections based on facial details, eye regions are often emphasized as the gaze direction of eyes are closely related to normal / abnormal driving patterns. Among eyes-based detection methods, the percentage of eyelid closure over the pupil over time (PERCLOS) [32] is popular. Technically, the percentage of closed eye time per unit time is utilized and it can be explicitly represented in Equation (1).

$$PERCLOS = \frac{Eye\ closing\ time}{Detection\ period} \times 100\% \tag{1}$$

When the percentage of time that the eyes' closure reaches 70% or even higher, the driver is normally considered to be in an abnormal driving state [32]. Although the PERCLOS detection method has merits of being effective and efficient, there are unfortunately several serious problems with it. First, eyes of drivers with different physiques and habits vary largely. An extreme case is that some people even do not close their eyes when sleeping, making the false positive of PERCLOS inevitably high. Second, some tough challenges, including unexpected movements of the head, will lead to detection failures of eyes. The above situations are certainly not beneficial for fulfilling abnormal driving behavior detection based on eyes [33]. For detections based on steering wheel, they are similar towards detections based on human physiological signals. In this study, a single visible-based camera is utilized to record high-resolution videos of the driver, and the automatic abnormal driving behavior detection is realized based on those captured video clips via sophisticated deep learning techniques. In this way, shortcomings of undesired high-variance regarding sensors (i.e., in detections based on either physiological signals or the steering wheel) can be totally avoided.

B. RECENT DEVELOPMENTS IN DEEP LEARNING AND ITS POPULAR UTILIZATIONS

It is interesting to notice that, deep learning techniques receive vast popularity when powerful computational hardware and large-scale data become more and more available nowadays. Generally speaking, most contemporary deep learning models can be categorized into two types, i.e., deep generative learning models and deep discriminant learning models. To be specific, deep generative learning models mainly aim to replicate ''fake-but-realistic'' data based on real data, and popular deep generative learning models include but not limited to VAE (i.e., variational auto-encoder) [34], GAN (i.e., generative adversarial network) [35], GLOW (i.e., generative flow) [36], etc. Deep discriminative learning models, on the other hand, are mainly utilized for discrimination / classification purposes. Well-known deep discriminative learning models are often winners of noticeable worldwide vision-based competitions (e.g., ILSVRC, COCO, etc.). Typical deep discriminative learning models include but not limit to AlexNet [37], VGG [38], GoogleNet [39], ResNet, etc.

Recently, contemporary deep learning models demonstrate the following trends. First, more and more deep learning models become tremendously deep for guaranteeing outstanding generalization capabilities. Second, their model structures become more and more sophisticated. For instance, the width of many contemporary deep learning models increases significantly. In [40], it is reported that a wide 40-layer ResNet model can gain the similar generalization capability as a conventional ''narrow'' single-channel 1001-layer ResNet model, but the wide model only costs 1/8 training time of the narrow one. Also, the cardinality of deep learning models increases greatly as well. Cardinality is often regarded as the number of isolated paths in a deep learning model, and those paths often share the same topological structure [41], making many contemporary deep learning models become multi-channel-based in their architectures. Therefore, important issues, including the depth, the wideness and the cardinality, are often carefully taken into consideration in designing structures of contemporary deep learning models. Furthermore, it is also necessary to point out that, other new deep learning model structures proposed in recent years, such as the capsule architecture in CapsNet [42], etc., also receive much popularity. In this study, three important issues mentioned above are incorporated in designing structures of the three newly introduced deep learning fusion models, for realizing the video-based abnormal driving behavior detection task.

Moreover, many up-to-date deep learning techniques have been vastly utilized in several vision-based utilizations in recent years. In the popular vision-based detection and tracking field, new and sophisticated architectures in deep learning have been proposed and utilized. In [43], a novel visual attention network was proposed to represent both global latent saliency and local latent saliency in the visual attention prediction task. In [44], a new hyper-parameter optimization method based on an action-prediction network leveraged by continuous deep Q-learning was introduced for object tracking. In [45], a novel triplet loss was proposed to extract expressive deep latent features from Siamese networks for fulfilling the same object tracking task. In [46], the deep metric learning technique based on a new multi-channel ResNet model was investigated for the tracking purpose. It is also interesting to point out that, concepts such as saliency, attentions, etc., are quite popular in recent video- / image-based detection utilizations [47], [48]. In this study, since the video-based abnormal driving behavior detection problem is

normally considered as a multi-class classification problem and solved mainly based on the global image plane, the idea of attention-aware deep learning is not incorporated. However, this idea of attention-aware deep learning is important and it can bring about ''global + local'' latent features-based abnormal driving behavior detection studies in the future. For other vision-based domains, new and sophisticated deep learning models are also thoroughly investigated. In [49], a new unbalanced deep discriminant learning model was proposed to fulfill the important medical images synthesis task in clinical diagnosis. In [50] and [51], the notion of disease similarity among multiple patients was emphasized and its learning can also be incorporated into deep learning models and realized through the classic back-propagated end-to-end learning procedure. It can be concluded based on the above descriptions that, deep learning models are widely and frequently utilized in various vision-based studies. Novel architectures of deep learning models as well as their associated new learning techniques are worthy of thorough investigated. In this study, novel architectures inspired by DenseNet as well as ResNet with superpositions of previous layers will be incorporated within newly proposed models for fulfilling the video-based abnormal driving behavior detection task.

III. METHODOLOGY

In this section, technical details of the three new deep learning fusion models for automatically detecting abnormal driving behavior are elaborated. Since the three new fusion models are inspired by DenseNet, the very model as well as other conventional deep learning models will be introduced in Section III-A first. It is worthy to mention that, all these conventional and popular deep learning models will be implemented and compared with three new deep learning-based fusion models in this study. After introducing conventional deep learning models in Section III-A, the three new deep learning-based fusion models will be elaborated in Section III-B. The energy function to be optimized in deep learning-based fusion models will be introduced in Section III-C.

A. CONVENTIONAL DEEP LEARNING MODELS

In the following, five conventional and popular deep learning models, including convolutional neural network (CNN), wide convolutional neural network (Wide CNN), group convolutional neural network (Group CNN), deep residual network (ResNet), and densely connected convolutional network (DenseNet), are introduced one by one. Details of their model structures utilized in this study are emphasized.

1) CONVOLUTIONAL NEURAL NETWORK (CNN)

One of the most earliest CNN models, i.e., the LeNet-5, was originally proposed for fulfilling the recognition and classification of handwritten characters, and its accuracy is satisfactory [52]. Generally speaking, the main architecture of CNN consists of the convolutional layer, the pooling layer, and the full connected layer. To be specific, the convolutional layer and the pooling layer work together to form multiple convolution groups, in order to extract latent features through a layer-by-layer model architecture. Then, the classification task can be completed based on latent features via fully connected layers. In this study, the model structure of CNN is depicted in Figure 1.

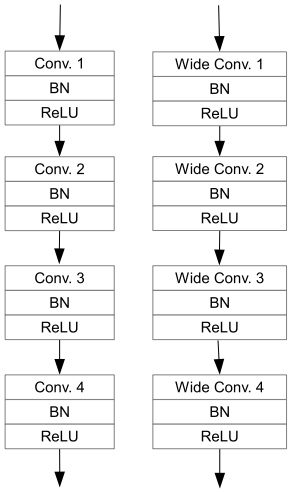

FIGURE 1. An illustration of model architectures in CNN (left) and Wide CNN (right) in this study.

2) WIDE CONVOLUTIONAL NEURAL NETWORK (WIDE CNN)

The idea of Wide CNN actually comes from the wide residual network (WRN) [40], which is on the basis of the deep residual network but further increases the number of layer-based convolution kernels. Figure 1 demonstrates the difference between the conventional CNN model and the Wide CNN model utilized in this study. It can be observed that, one significant difference between CNN and Wide CNN is that, wide convolution layers instead of the traditional ''narrow'' convolution layers are incorporated in Wide CNN. The motivation can be explained as follows. It is widely acknowledged that, it is challenging for gradients to be back-propagated when a deep learning model becomes tremendously deep, and such a tremendously deep model is often hard to be comprehensively trained. In order to tackle the above dilemma, WRN with a shallow but significantly wider architecture is proposed [40]. It is encouraging because, the generalization capability of this shallow but wide architecture outperforms that of the conventional deep and narrow architecture. Also, the former is easier to be thoroughly trained. In this study, Wide CNN is also implemented for experimental evaluations.

3) GROUP CONVOLUTIONAL NEURAL NETWORK (GROUP CNN)

Group CNN mainly counts on group convolutions, which are quite different from traditional convolutions adopted in the vast majority of CNN-based deep learning models. To be specific, each individual convolution filter in CNN operates on all channels, while each individual convolution filter in

Group CNN is active only on partial channels. An illustration of the difference between traditional convolution filters in CNN and group convolution filters in Group CNN is depicted in Figure 2. It can be noticed that, Figure 2 describes the case of 2-channel. For traditional convolution filters in CNN (i.e., left in Figure 2), C traditional convolution filters are executed on N feature maps, in order to obtain C feature maps. For group convolution filters in Group CNN (i.e., right in Figure 2), N feature maps are divided into two parts (i.e., each part contains $\frac{N}{2}$ feature maps for balanced considerations). Each individual part is then fed into $\frac{C}{2}$ convolution filters, in order to generate $\frac{C}{2}$ feature maps. In this way, different convolution filters are actually executed on different channels. The idea of Group CNN is important, as different feature maps can be generated using different GPUs and the final result can be fused based on them, making the model more efficient to be trained with multiple GPUs. Additionally, the recently popular ResNeXt model also adopts the above group convolution idea in its residual network-based architecture [41].

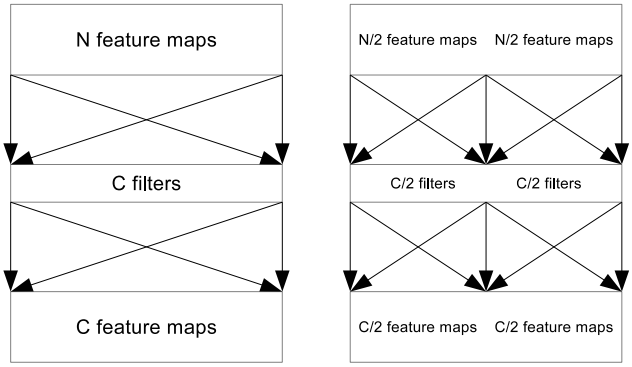

FIGURE 2. An illustration of the difference between traditional convolution filters in CNN (left) and group convolution filters in Group CNN (right).

4) DEEP RESIDUAL NETWORK (ResNet)

ResNet is widely acknowledged as one of the most influential deep discriminant learning-based models at the current stage. ResNet is quite successful, as it is efficient to tackle the notorious problem of vanishing gradients that becomes commonly seen in many tremendously deep models. The core residual architecture in ResNet is shown in Figure 3, and its main idea is to add a parallel identity mapping to the original network, which is helpful to constitute a residual learning structure. Technically, provided the potential nonlinear mapping which needs to be learned as $H(x)$, a nonlinear stacked network can be constructed to represent the residual mapping $F(x) = H(x) - x$. In this way, the potential nonlinear mapping to be learned can be written as $F(x) + x$. It is also necessary to point out that, it is often easier to optimize the residual mapping $F(x)$ than to directly optimize the potential nonlinear mapping $H(x)$, which makes ResNet quite valuable.

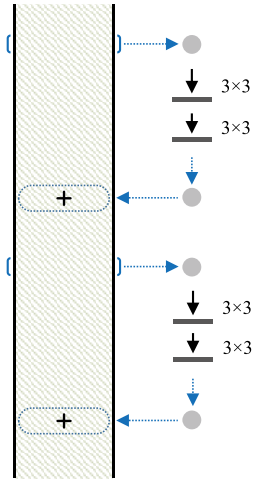

FIGURE 3. An illustration of the core residual architecture in ResNet.

5) DENSELY CONNECTED CONVOLUTIONAL NETWORK (DENSEnET)

DenseNet is the basic deep learning model inspiring new three deep learning-based fusion models in this study. Generally speaking, several well-established models belong to the big family of DenseNet, which includes Highway Network [53], GoogleNet, etc. Compared with ResNet, DenseNet is more thoroughgoing. The reason is because that, ResNet only adds outputs of two adjacent layers as illustrated in Figure 3, while DenseNet needs to add the current layer to all its previous layers (i.e., as illustrated in Figure 4). For instance, when there are L layers, ResNet prones to have $(L-1)$ direct connections (i.e., one direct connection appears between two adjacent layers). However, DenseNet will obtain $\binom{2}{L+1} = \frac{L\times(L+1)}{2}$ connections, totally. The advantage of DenseNet is that, gradient back-propagated transmission in DenseNet actually changes from the ''linear-like'' flow (i.e., in most conventional deep learning models) to the new ''tree-like'' flow, and the potential possibility of gradient vanishing will be greatly reduced in DenseNet. Also, the training efficiency of DenseNet will be boosted, therein.

B. THREE NOVEL DEEP LEARNING-BASED FUSION MODELS: WGD, WGRD, AND AWGRD

In the following, three novel deep learning-based fusion models inspired by DenseNet are introduced to tackle the video-based abnormal driving behavior detection problem for the first time. The three new models are named as the wide group dense (WGD) network, the wide group residual dense (WGRD) network, and the alternative wide group residual dense (AWGRD) network, respectively. Technically, WGD takes important issues of deep learning models, i.e., the depth, the width and the cardinality, into consideration when designing its model structure based on DenseNet. For WGRD and AWGRD, they are more sophisticated as the important idea of residual networks with superpositions of previous layers is incorporated. Technical details are as follows.

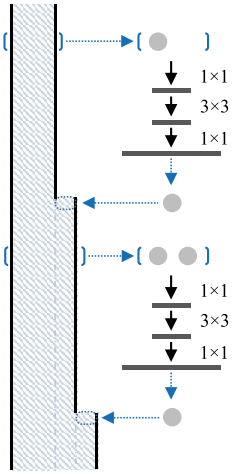

FIGURE 4. An illustration of the tree-like connection architecture in DenseNet.

1) WIDE GROUP DENSELY NETWORK (WGD)

The model architecture of WGD is demonstrated in Figure 5. It can be noticed that, the conventional convolution in DenseNet is replace by the group and wide convolution in WGD. The merit is that, the generalization capability of WGD can be improved via group and wide convolutions in WGD, while the number of parameters in WGD will not increase much. Also, the important enhancement of width and cardinality in WGD can be realized, therein.

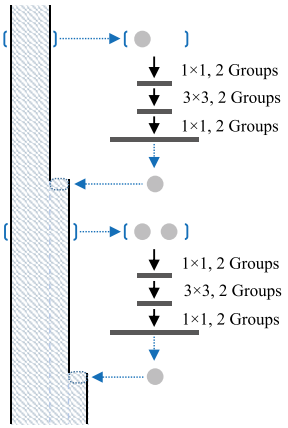

FIGURE 5. The model architecture of WGD.

2) WIDE GROUP RESIDUAL DENSELY NETWORK (WGRD)

The model architecture of WGRD is depicted in Figure 6. The most significant change of WGRD with respect to WGD is that, the idea of residual networks is incorporated in WGRD. Details of WGRD and its affinity with other deep learning models can be explained as follows. Provided an input image x_0 transmitted through a L-layer network, and the l-layer can be represented via a non-linear transformation $H_l(\cdot)$ (i.e., composed of BN, ReLU, Conv, etc.). Let the output of the l-layer be x_l. Then, x_l in a conventional feedforward network

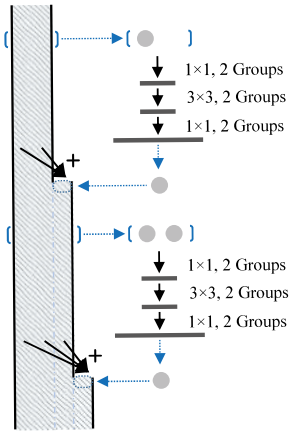

FIGURE 6. The model architecture of WGRD.

can be represented in Equation 2.

$$x_l = H_l(x_{l-1}) \tag{2}$$

For x_l in a typical ResNet architecture, it can be obtained as the addition between the input and the output of the l-layer. Therefore, x_l of ResNet can be described in Equation 3.

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{3}$$

For DenseNet and WGD, the situation becomes more sophisticated. Since ResNet only takes the input and the output of the l-layer into consideration, intermediate outputs from other previous layers (i.e., x_1 from the $1st$-layer, x_2 from the $2nd$-layer, $\ldots$, x_{l-2} from the $(l-2)$-layer) will be totally neglected. In order to tackle the above problem, DenseNet and WGD take all above-mentioned information into consideration when representing x_l, which can be represented in Equation 4 (i.e., the operator $[\cdot]$ represents the parallel operation).

$$x_l = H_l[x_0, x_1, \cdots, x_{l-1}] \tag{4}$$

$$x_l = H_l[x_0, x_0 + x_1, \cdots, \sum_{i=0}^{l-1} x_i] \tag{5}$$

For WGRD, a more sophisticated constitution of x_l is further applied. As described in Equation 5, the superposition of previous layers (i.e., $\sum x_i$) is utilized on the i-layer. The reason is because that, more complex features are added as the input of the l-layer, and the learning capability of the network can be strengthen, therein. In this way, the important idea of residual networks with superpositions of previous layers can be realized in WGRD.

3) ALTERNATIVE WIDE GROUP RESIDUAL DENSELY NETWORK (AWGRD)

In this study, an alternative WGRD (i.e., AWGRD) is also introduced to fulfill the video-based abnormal driving behavior detection task. The model architecture of AWGRD is illustrated in Figure 7, and its main idea is described in

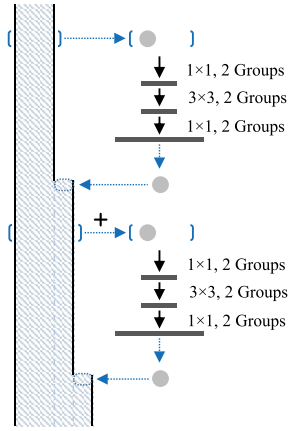

FIGURE 7. The model architecture of AWGRD.

Equation 6.

$$x_l = H_l[\sum_{i=0}^{l-1} x_i] \tag{6}$$

It is easy to notice that, x_l in Equation 6 only takes the superposition of previous $l-1$ layers (i.e., $\sum_{i=0}^{l-1} x_i$) into consideration, while x_l of WGRD in Equation 5 takes superposition of all previous layers (i.e., $x_0, x_0+x_1, \cdots, \sum_{i=0}^{l-1} x_i$) into consideration. Therefore, AWGRD can be regarded as a simplified version of WGRD, but its training efficiency will undoubtedly becomes higher. The generalization capability of AWGRD in automatically detecting abnormal driving behavior will be quantitatively demonstrated in Section IV, from the statistical perspective.

C. THE ENERGY FUNCTION TO BE OPTIMIZED IN DEEP LEARNING-BASED FUSION MODELS

The video-based abnormal driving behavior detection task in this study can be regarded as a general multi-class classification problem, in which the classic cross entropy is utilized to constitute the energy function to be optimized. Provided the i-th image (i.e., frame) of a video clip as x_i, and its label information as y_i (i.e., y_i is represented via a c-dimensional feature vector in this study, while c indicates the number of classes). Let $y_i' = P(x_i)$ denote the probability that a deep learning-based fusion model assigns x_i to one particular class (i.e., $P(\cdot)$ indicates the whole mapping of the deep learning-based fusion model), and the energy function based on cross entropy in the three deep learning-based fusion models can be represented in Equation 7.

$$L = -\sum_{i=1}^{m}\sum_{j=1}^{c} y_{ij} log(y_{ij}') = -\sum_{i=1}^{m}\sum_{j=1}^{c} y_{ij} log P(x_{ij}) \tag{7}$$

where, m represents the number of images. It is easy to perceive that, Equation 7 aims to minimize the difference between the predicted probability distribution (i.e., y_i') and the real probability distribution (i.e., y_i). The whole optimization can be realized via the conventional stochastic gradient descent (SGD) algorithm in this study.

When trainings of deep learning-based fusion models are complete, such learned models will be utilized to automatically detect abnormal driving behavior in real time. The pseudo-code including main steps to fulfill this real-time video-based abnormal driving behavior detection in this study is elaborated in Algorithm 1. It can be perceived that, real-time video as well as a learned deep learning model are both required as inputs in Algorithm 1. Then, for a video clip, each individual frame of it will be extracted and then to be tested via the learned deep learning model. The model aims to classify the driving pattern within the input image, and if the pattern belongs to ones of abnormal drivings, a warning message will be sent to the driver, immediately.

Algorithm 1 The Pseudo-Code to Fulfill the Real-Time Video-Based Abnormal Driving Behavior Detection in This Study

Require: V: Real-time Video M: A Learned Deep Learning Model
for all frame in V **do**
  Feed each individual frame into M for determining its driving pattern;
  if pattern belongs to ones of abnormal drivings **then**
    warn the driver;
  end if
end for

IV. EXPERIMENTS

A. DATABASE AND EXPERIMENTAL SETTINGS

To verify the effectiveness of newly proposed deep learning-based fusion models in automatically detecting abnormal driving behavior of this study, the Kaggle state farm distracted driver detection database was utilized [54]. To be specific, there are totally 22,424 color frames (i.e., images) of drivers in video clips of this database. Each individual image has a fixed spatial resolution of 640×480, and all images can be categorized into 10 classes, which indicates 10 different driving patterns. These driving patterns contain safe driving, texting (using right hand), talking on the phone (using right hand), texting (using left hand), talking on the phone (using left hand), operating the radio, drinking, reaching behind, hair and makeup, talking to passenger, etc. Example images of them are displayed in Figure 8.

All deep learning models introduced in Section III are implemented for comparisons in this experiment. In order to make the database fit sophisticated deep learning models better, the data augmentation is realized. The above step is fulfilled by a series of image pre-processing, which include noise additions, intensity changes, color changes, image rotations, image scaling, etc. After executing the above step, all images in the database are randomly and evenly divided into two parts, i.e., the training database and the testing database. For implementations of deep learning models, the batch size

FIGURE 8. Example images of 10 driving patterns in the utilized Kaggle state farm distracted driver detection database (from left to right, up to bottom: Safe driving, texting (right hand), talking on the phone (right hand), texting (left hand), talking on the phone (left hand), operating the radio, drinking, reaching behind, hair and makeup, talking to passenger).

is 32, the number of training epochs is 10, and the learning rate is 0.0001. The above parameters are pre-defined after fulfilling trails-and-errors for optimal detection performance. For parameters to be learned within all deep learning models, their detailed numbers are elaborated in Table 1. It can be observed that, three new deep learning-based fusion models have less parameters compared with most conventional deep learning models (e.g., CNN, Wide CNN, Group CNN, ResNet, etc.), so that training efficiencies of new deep learning-based fusion models can be favored. All trainings are implemented using a workstation equipped with Intel Xeon Silver 4110 CPU, 128G RAM, Nvidia Titan V GPU card, CentOS 7 and PyTorch 1.0.0.

TABLE 1. Numbers of parameters to be learned in all deep learning models of this study.

Model	Parameters	Model	Parameters
CNN	9.95M	Wide CNN	19.36M
Group CNN	5.25M	ResNet	11.18M
DenseNet	0.41M	WGD	1.54M
WGRD	1.58M	AWGRD	1.42M

B. EXPERIMENTAL RESULTS AND STATISTICAL ANALYSES

Figure 9 demonstrates the trend of accuracies increasing with respect of training epochs in all compared deep learning models. First, it can be noticed that, accuracies of all deep learning models keep on increasing and then become stable when their training epochs further increase, which is a significant indicator of the thorough training and convergence of all deep learning models. Second, three deep learning-based fusion models, DenseNet, as well as ResNet outperform other conventional CNN-based models (i.e., CNN, Wide CNN, Group CNN) as revealed in Figure 9. For comparisons between three deep learning-based fusion models and DenseNet, it is interesting to notice that, the former reaches the stable stage faster (i.e., less epochs) than DenseNet, and significant robustness can be obtained from new deep learning-based fusion models.

In order to quantitatively evaluate the detection accuracy from statistical point of view, the classic precision-recall (P-R) curve is utilized. Figure 10 demonstrates P-R curves of all deep learning models in this study. It is necessary to point out that, the area under a P-R curve indicates the mean average precision (MAP) of the corresponding model.

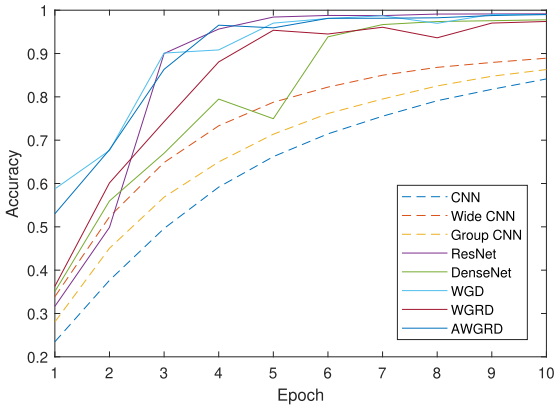

FIGURE 9. The trend of accuracies increasing with respect of training epochs in all deep learning models.

It can be observed from Figure 10 that, AWGRD achieves the highest MAP among all compared models. For conventional CNN-based models (i.e., CNN, Wide CNN, Group CNN), their P-R curves are significantly lower than those of others, which indicates that more sophisticated model architectures (e.g., ResNet-based, DenseNet-based, etc.) are beneficial for correctly detecting abnormal driving behavior in this study. Another interesting observation from Figure 10 is that, DenseNet and its derivatives (i.e., three novel deep learning-based fusion models) outperform ResNet regarding their P-R curves, which suggests that the superposition of previous layers in DenseNet is superior to incorporating only one previous layer in ResNet for automatically detecting abnormal driving behavior in this study.

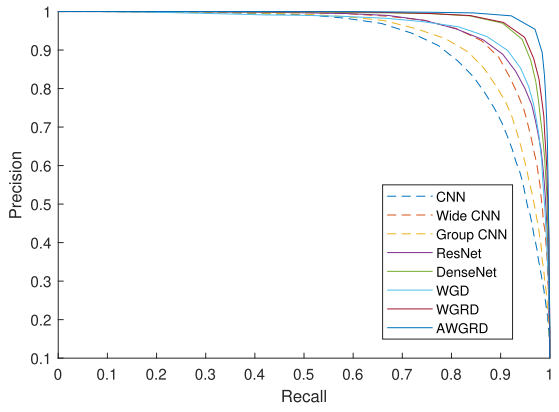

FIGURE 10. Precision-recall curves of all deep learning models in this study.

Precision and recall outcomes are then utilized to further calculate the unbiased F-measure (i.e., $F-measure = \frac{2\times precision\times recall}{precision+recall}$), and the box-and-whisker plot of F-measures calculated from precision and recall outcomes of all deep learning models are illustrated in Figure 11. In each individual box of Figure 11, the red horizontal line in each box represents the median of F-measure, while the upper and lower quartiles of F-measure is represented by blue lines

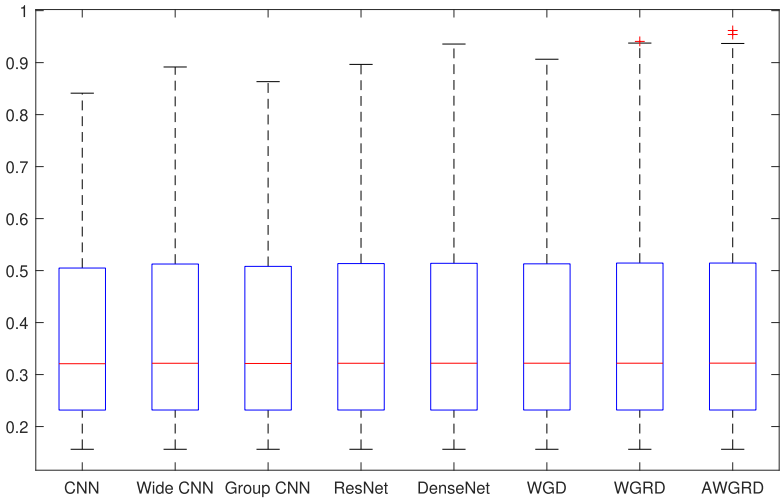

FIGURE 11. The box-and-whisker plot of F-measures calculated from precision and recall outcomes of all deep learning models.

above and below the median in each box. A vertical dashed line is drawn from the upper quartile and the lower quartile to their most extreme data points, which are within the 1.5 inter-quartile range (IQR). Each individual data beyond the 1.5 IQR is marked via a plus sign. Furthermore, a more detailed quantitative analysis made up of one-way analysis of variance (ANOVA) and multiple comparison tests are conducted based on unbiased F-measure outcomes. In one-way ANOVA, F-measure results obtained from all deep learning models are compared to test a hypothesis (H_0) that ''F-measure means of all deep learning models are equivalent'', against the general alternative that these means cannot be all the same. The p-value is utilized here as an indicator to reveal whether H_0 holds or not. In this study, p-values calculated from all F-measure results are nearly 0, which is a strong indication that all these models cannot share the same F-measure mean. Therefore, the next step is to conduct more detailed paired comparisons. The reason to do so is because that, the alternative against H_0 is too general. Information about which deep learning model is superior from the statistical perspective cannot be perceived by one-way ANOVA alone. There are two kinds of evaluation after applying multiple comparison tests on calculated F-measure of all models, and quantitative evaluation results are shown in Tables 2, 3, and 4. For the two kinds of evaluation, one is estimated F-measure mean difference, which is a single-value estimator of F-measure mean difference. Another is a 95 % confidence interval (CI). In statistics, a CI is a special form of interval estimator for a parameter (i.e. F-measure mean difference in this experiment). Generally speaking, instead of estimating the parameter by a single value, CI is capable to provide an interval estimation which is likely to include the estimated parameter within a specified interval.

It can be summarized from Tables 2, 3, and 4 that, the majorities of F-measure entries are positive. Since the three deep learning-based fusion models are utilized as Method 1 in above-mentioned three tables (i.e., WGD in Table 2, WGRD in Table 3, and AWGRD in Table 4), those positive entries in tables substantiate the superiority of Method 1 to other compared conventional deep learning models. It is also necessary to point out that, for the comparison between WGD and DenseNet-29 in Table 2, the single-value estimation is negative and its corresponding 95% confidence interval is also negative. It suggests that DenseNet-29 outperforms WGD based on F-measure in this study, which complies well with the fact that the P-R curve of DenseNet is above that of WGD in Figure 10.

TABLE 2. Multiple comparison tests between WGD and other conventional models based on F-measure in this study.

Model 1	Model 2	F-measure Mean Difference	A 95% Confidence Interval
WGD	CNN-12	0.00099	[0, 0.06529]
WGD	Wide CNN-12	0.00012	[0, 0.01491]
WGD	Group CNN-12	0.00052	[0, 0.04313]
WGD	ResNet-18	0.00012	[0, 0.00992]
WGD	DenseNet-29	-0.00005	[-0.02922, 0]

TABLE 3. Multiple comparison tests between WGRD and other conventional models based on F-measure in this study.

Model 1	Model 2	F-measure Mean Difference	A 95% Confidence Interval
WGRD	CNN-12	0.00099	[0, 0.09960]
WGRD	Wide CNN-12	0.00012	[0, 0.04922]
WGRD	Group CNN-12	0.00052	[0, 0.07744]
WGRD	ResNet-18	0.00012	[0, 0.04423]
WGRD	DenseNet-29	0.00005	[0, 0.00509]

TABLE 4. Multiple comparison tests between AWGRD and other conventional models based on F-measure in this study.

Model 1	Model 2	F-measure Mean Difference	A 95% Confidence Interval
AWGRD	CNN-12	0.00110	[0, 0.12057]
AWGRD	Wide CNN-12	0.00023	[0, 0.07019]
AWGRD	Group CNN-12	0.00063	[0, 0.09841]
AWGRD	ResNet-18	0.00023	[0, 0.06520]
AWGRD	DenseNet-29	0.00016	[0, 0.02606]

Another detailed comparison is carried out regarding the three new deep learning-based fusion models, and statistical outcomes are demonstrated in Table 5. It can be concluded that, AWGRD outperforms WGD and WGRD based on F-measure in this study. Since AWGRD also incorporates the important idea of residual networks with superpositions of previous layers but with less parameters (i.e., 1.42M parameters compared with 1.54M parameters of WGD and 1.58M parameters of WGRD as indicated in Table 1), it finds a good balance between the effectiveness and efficiency among all three newly proposed models in this study.

TABLE 5. Multiple comparison tests among AWGRD, WGD and WGRD based on F-measure in this study.

Model 1	Model 2	F-measure Mean Difference	A 95% Confidence Interval
AWGRD	WGD	0.00011	[0, 0.05528]
AWGRD	WGRD	0.00011	[0, 0.02097]

C. DISCUSSIONS

Experimental analyses in Section IV-B demonstrate the superiority of newly introduced deep learning-based fusion models in automatically detecting abnormal driving behavior from the statistical point of view. In this section, more details about situations in which new models can outperform conventional deep learning models will be discussed.

According to Section IV-B, WGD can outperform ResNet, and example cases in which WGD correctly classifies driving patterns but ResNet cannot do are displayed in Figure 12. It can be observed that, when the lighting condition within the vehicle is not ample, ResNet becomes more prone to misclassify visually similar driving patterns. However, WGD is more capable to discern the correct driving pattern in this challenging situation. For Figures 13 and 14, example cases which reveal that WGRD outperforms ResNet and DenseNet are displayed, respectively. For Figures 15 and 16, additional example cases suggesting that AWGRD outperforms ResNet and DenseNet are displayed, respectively. It can be concluded based on these example cases that, WGRD and AWGRD are capable to correctly discern visually similar cases, especially within challenging conditions (e.g., poor lighting, partial occlusions of cell phones, etc.).

Another interesting discussion is as follows. Three newly introduced deep learning-based fusion models are executed on a Nvidia Titan V GPU card, in order to evaluate their efficiency when detecting different driving patterns (i.e., at the testing stage). Table 6 elaborates their operation times of correctly detecting different driving patterns. It is quite encouraging that, the average operation times of all three new models are around 30 *ms*. It is promising to realize real-time operations of automatic driving patterns detection using the three newly introduced deep learning-based fusion models based on one single visible-based camera and a Nvidia Titan V GPU card (i.e., 30 *ms* can guarantee no less than 33 *fps* that meets the requirement of real-time operations).

FIGURE 12. Example cases in which WGD correctly classifies driving patterns but ResNet cannot do.

FIGURE 13. Example cases in which WGRD correctly classifies driving patterns but ResNet cannot do.

FIGURE 14. Example cases in which WGD correctly classifies driving patterns but DenseNet cannot do.

FIGURE 15. Example cases in which AWGD correctly classifies driving patterns but ResNet cannot do.

FIGURE 16. Example cases in which AWGD correctly classifies driving patterns but DenseNet cannot do.

TABLE 6. Operation times (mean ± standard deviation) of correctly detecting different driving patterns via WGD, WGRD, and AWGRD in this study (units: *ms*).

Driving Patterns	WGD	WGRD	AWGRD
safe driving	27.3 ± 17.4	26.7 ± 16.8	33.1 ± 23.1
texting (right hand)	27.4 ± 17.5	26.5 ± 16.6	33.4 ± 22.5
talking on phone (right hand)	27.0 ± 16.1	26.7 ± 17.1	32.7 ± 22.2
texting (left hand)	27.7 ± 17.7	26.4 ± 16.5	32.9 ± 22.3
talking on phone (left hand)	27.5 ± 17.7	26.3 ± 15.9	32.2 ± 22.9
operating the radio	27.0 ± 16.7	27.4 ± 16.9	33.2 ± 23.4
drinking	27.6 ± 17.6	26.3 ± 16.6	33.1 ± 22.7
reaching behind	27.5 ± 16.8	25.6 ± 15.1	31.8 ± 20.5
hair and makeup	26.2 ± 16.0	26.7 ± 16.8	31.8 ± 21.2
talking to passenger	27.3 ± 17.6	25.5 ± 15.5	32.1 ± 22.3
average operation times	27.3 ± 17.1	26.4 ± 16.4	32.7 ± 22.4

V. CONCLUSION

The video-based abnormal driving behavior detection study is highly important nowadays, as it is a reliable and automatic manner to ensure safeties of drivers. Also, it receives vast popularity as it is an essential step to realize fully automatic driving (i.e., particularly in Level-3 and Level-4 stages according to the ''autonomous driving'' definition provided by the US Department of Transportation's National Highway Traffic Safety Administration). In this study, three novel deep learning-based fusion models are introduced for the first time, to fulfill the video-based abnormal driving behavior detection task. Technically, these new models are inspired by the popular DenseNet, which was proposed in recent years. For WGD, it emphasizes on important issues of designs of modern deep learning models, including the depth, the width, and the cardinality. The width and the cardinality of WGD significantly increase, therein. For WGRD and AWGRD, they are more sophisticated as the important idea of residual networks with superpositions of previous layers is incorporated. This idea is highly valuable in the video-based abnormal driving behavior detection task, as temporary and spatial latent information can be comprehensively described with the help of superpositions of previous layers. Extensive experiments based on the standard Kaggle state farm distracted driver detection dataset as well as rigorous comparisons with several other popular deep learning models suggest the superiority of newly proposed deep learning-based fusion models in both effectiveness and efficiency. In the future, ''global + local'' latent features-based abnormal driving behavior detection studies will be conducted. Also, effective and efficient deep learning models realized via customized mobile chips will be investigated for realizing abnormal driving behavior detection.

REFERENCES

[1] W. Cao, J. Yuan, Z. He, Z. Zhang, and Z. He, ''Fast deep neural networks with knowledge guided training and predicted regions of interests for real-time video object detection,'' *IEEE Access*, vol. 6, pp. 8990–8999, 2018.

[2] H. Shuai, Q. Liu, K. Zhang, J. Yang, and J. Deng, ''Cascaded regional spatio-temporal feature-routing networks for video object detection,'' *IEEE Access*, vol. 6, pp. 3096–3106, 2018.

[3] A. Nanda, P. K. Sa, S. K. Choudhury, S. Bakshi, and B. Majhi, ''A neuromorphic person re-identification framework for video surveillance,'' *IEEE Access*, vol. 5, pp. 6471–6482, 2017.

[4] L. Sun, Z. Jiang, H. Song, Q. Lu, and A. Men, ''Semi-coupled dictionary learning with relaxation label space transformation for video-based person re-identification,'' *IEEE Access*, vol. 6, pp. 12587–12597, 2018.

[5] Y. Wu, Y. Sui, and G. Wang, ''Vision-based real-time aerial object localization and tracking for UAV sensing system,'' *IEEE Access*, vol. 5, pp. 23969–23978, 2017.

[6] S.-H. Lee, M.-Y. Kim, and S.-H. Bae, ''Learning discriminative appearance models for online multi-object tracking with appearance discriminability measures,'' *IEEE Access*, vol. 6, pp. 67316–67328, 2018.

[7] M. S. Hossain and G. Muhammad, ''An emotion recognition system for mobile applications,'' *IEEE Access*, vol. 5, pp. 2281–2287, 2017.

[8] Z. Pan, X. Yi, and L. Chen, ''Motion and disparity vectors early determination for texture video in 3D-HEVC,'' *Multimedia Tools Appl.*, to be published. doi: 10.1007/s11042-018-6830-7.

[9] J. Wang, Z. Zhang, B. Li, S. Lee, and R. S. Sherratt, ''An enhanced fall detection system for elderly person monitoring using consumer home networks,'' *IEEE Trans. Consum. Electron.*, vol. 60, no. 1, pp. 23–29, Feb. 2014.

[10] Z. Zhang, X. Guo, and Y. Lin, ''Trust management method of D2D communication based on RF fingerprint identification,'' *IEEE Access*, vol. 6, pp. 66082–66087, 2018.

[11] J. Wang, Y. Cao, B. Li, H.-J. Kim, and S. Lee, ''Particle swarm optimization based clustering algorithm with mobile sink for WSNs,'' *Future Gener. Comput. Syst.*, vol. 76, pp. 452–457, Nov. 2017.

[12] Z. Xue, J. Wang, G. Ding, Q. Wu, Y. Lin, and T. A. Tsiftsis, ''Device-to-device communications underlying UAV-supported social networking,'' *IEEE Access*, vol. 6, pp. 34488–34502, 2018.

[13] J. Wang, J. Cao, R. S. Sherratt, and J. H. Park, ''An improved ant colony optimization-based approach with mobile sink for wireless sensor networks,'' *J. Supercomput.*, vol. 74, no. 12, pp. 6633–6645, Dec. 2018.

[14] Y. Tu, Y. Lin, J. Wang, and J.-U. Kim, ''Semi-supervised learning with generative adversarial networks on digital signal modulation classification,'' *Comput. Mater. Continua*, vol. 55, no. 2, pp. 243–254, 2018.

[15] J. Wang, Y. Gao, X. Yin, F. Li, and H.-J. Kim, ''An enhanced PEGASIS algorithm with mobile sink support for wireless sensor networks,'' *Wireless Commun. Mobile Comput.*, vol. 2018, Dec. 2018, Art. no. 9472075.

[16] J. Sun *et al.*, ''A multi-focus image fusion algorithm in 5G communications,'' *Multimedia Tools Appl.*, vol. 3, pp. 1–20, Feb. 2018.

[17] J. Wang, C. Ju, Y. Gao, A. K. Sangaiah, and G.-J. Kim, ''A PSO based energy efficient coverage control algorithm for wireless sensor networks,'' *Comput. Mater. Continua*, vol. 56, no. 3, pp. 433–446, Sep. 2018.

[18] Y. Lin, X. Zhu, Z. Zheng, Z. Dou, and R. Zhou, ''The individual identification method of wireless device based on dimensionality reduction and machine learning,'' *J. Supercomput.*, vol. 5, pp. 1–18, Dec. 2017.

[19] E. B. Tirkolaee, A. A. R. Hosseinabadi, M. Soltani, A. K. Sangaiah, and J. Wang, ''A hybrid genetic algorithm for multi-trip green capacitated arc routing problem in the scope of urban services,'' *Sustainability*, vol. 10, no. 5, p. 1366, 2018.

[20] W. Qidi, L. Yibing, L. Yun, and Y. Xiaodong, ''The nonlocal sparse reconstruction algorithm by similarity measurement with shearlet feature vector,'' *Math. Problems Eng.*, vol. 2014, Mar. 2014, Art. no. 586014.

[21] Q. Wu, Y. Li, and Y. Lin, ''The application of nonlocal total variation in image denoising for mobile transmission,'' *Multimedia Tools Appl.*, vol. 76, no. 16, pp. 17179–17191, Aug. 2016.

[22] D. Zhang, D. Meng, and J. Han, ''Co-saliency detection via a self-paced multiple-instance learning framework,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[23] D. Zeng, Y. Dai, F. Li, R. S. Sherratt, and J. Wang, ''Adversarial learning for distant supervised relation extraction,'' *Comput. Mater. Continua*, vol. 55, no. 1, pp. 121–136, 2018.

[24] R. Olson, R. Hanowski, J. Hickman, and J. Bocanegra, ''Driver distraction in commercial vehicle operations,'' U.S. Dept. Transp., Rep. Federal Motor Carrier Saf. Admin., Washinton, DC, USA, Tech. Rep. FMCSA-RRT-09-042, 2009.

[25] G. Huang, Z. Liu, L. Van der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.

[26] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[27] A. Saeed, S. Trajanovski, M. Van Keulen, and J. Van Erp, ''Deep physiological arousal detection in a driving simulator using wearable sensors,'' in *Proc. ICDMW*, Nov. 2017, pp. 486–493.

[28] Z. Pan, H. Qin, X. Yi, Y. Zheng, and A. Khan, ''Low complexity versatile video coding for traffic surveillance system,'' *Int. J. Sensor Netw.*, vol. 30, no. 2, pp. 116–125, 2019.

[29] M. Jeong, B. C. Ko, S. Kwak, and J.-Y. Nam, ''Driver facial landmark detection in real driving situations,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2753–2767, Oct. 2018.

[30] V. Balasubramanian and R. Bhardwaj, ''Grip and electrophysiological sensor-based estimation of muscle fatigue while holding steering wheel in different positions,'' *IEEE Sensors J.*, vol. 19, no. 5, pp. 1951–1960, Mar. 2019.

[31] N. Li, T. Misu, and F. Tao, ''Understand driver awareness through brake behavior analysis: Reactive versus intended hard brake,'' in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1523–1528.

[32] W. Wierwille and R. Knipling, ''Vehicle-based drowsy driver detection: Current status and future prospects,'' in *Proc. IVHS*, 1994, pp. 1–24.

[33] A. Acioğlu and E. Erçelebi, ''Real time eye detection algorithm for PERCLOS calculation,'' in *Proc. 24th Signal Process. Commun. Appl. Conf.*, pp. 1641–1644, May 2016.

[34] D. P. Kingma and M. Welling. (2013). *Auto-Encoding Variational Bayes*. [Online]. Available: https://arxiv.org/abs/1312.6114

[35] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
[36] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. NIPS*, 2018, pp. 10215–10224.
[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[38] K. Simonyan and A. Zisserman. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [Online]. Available: https://arxiv.org/abs/1409.1556
[39] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.
[40] S. Zagoruyko and N. Komodakis. (2017). *Wide Residual Networks*. [Online]. Available: https://arxiv.org/abs/1605.07146
[41] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. CVPR*, Jul. 2017, pp. 1492–1500.
[42] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. NIPS*, 2017, pp. 3856–3866.
[43] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.
[44] X. Dong, J. Shen, W. Wang, Y. Liu, L. Shao, and F. Porikli, "Hyperparameter optimization for tracking with continuous deep Q-learning," in *Proc. CVPR*, Jun. 2018, pp. 518–527.
[45] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proc. ECCV*, Sep. 2018, pp. 459–474.
[46] W. Huang, H. Ding, and G. Chen, "A novel deep multi-channel residual networks-based metric learning method for moving human localization in video surveillance," *Signal Process.*, vol. 142, pp. 104–113, Jan. 2018.
[47] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. doi: 10.1109/TPAMI.2018.2840724.
[48] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.
[49] W. Huang *et al.*, "Arterial spin labeling images synthesis from sMRI using unbalanced deep discriminant learning," *IEEE Trans. Med. Imag.*, to be published. doi: 10.1109/TMI.2019.2906677.
[50] W. Huang, S. Zeng, M. Wan, and G. Chen, "Medical media analytics via ranking and big learning: A multi-modality image-based disease severity prediction study," *Neurocomputing*, vol. 204, pp. 125–134, Sep. 2016.
[51] W. Huang, "A novel disease severity prediction scheme via big pair-wise ranking and learning techniques using image-based personal clinical data," *Signal Process.*, vol. 124, pp. 233–245, Jul. 2016.
[52] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[53] R. K. Srivastava, K. Greff, and J. Schmidhuber. (2015). *Highway Networks*. [Online]. Available: https://arxiv.org/abs/1505.00387
[54] *Kaggle State Farm Distracted Driver Detection Dataset*. Accessed: Apr. 25, 2019. [Online]. Available: https://www.kaggle.com/c/state-farm-distracted-driver-detection

WEI HUANG received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, China, in 2004 and 2006, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2011. He was with the University of California at San Diego, USA, and the Agency for Science Technology and Research, Singapore, as a Postdoctoral Research Fellow. He joined Nanchang University as a Faculty Member. He has published more than 70 academic papers and has been acting as a principal investigators in more than 10 national/provincial grants, including three NSFC grants and two NSF key grants in Jiangxi Province. His research interests include image processing, pattern recognition, machine learning, and computer vision. He received the Best Paper Award of MICCAI-MLMI, in 2010, the most interesting paper award of ICME-ASMMC, in 2016, and was entitled the provincial young scientist of Jiangxi Province, in 2015.

XI LIU received the B.Eng. degree from Nanchang University, in 2017, where she is currently pursuing the M.Eng. degree under the supervision of Prof. W. Huang. Her research interests include computer vision and pattern recognition.

MINGYUAN LUO received the B.Eng. degree from Nanchang University, in 2017, where he is currently pursuing the M.Eng. degree under the supervision of Prof. W. Huang. His research interests include medical image processing, machine learning, computer vision, and pattern recognition.

PENG ZHANG received the B.E. degree from Xi'an Jiaotong University, China, in 2001, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2011. He is currently a Professor with the School of Computer Science, Northwestern Polytechnical University, China. His current research interests include image processing, signal processing, and computer vision. He is also the TPC Member and a Reviewer of several high ranked international conferences and journals.

WEI WANG received the Ph.D. degree in the international business program with an emphasis in data mining and management from Texas A&M International University, USA. She is currently an Assistant Professor with Chang'an University, China. Her current research interests include data mining, machine learning, and related data-driven learning-based applications.

JIN WANG received the M.S. degree from the Nanjing University of Posts and Telecommunications, China, in 2005, and the Ph.D. degree from Kyung Hee University, South Korea, in 2010. He is currently a Professor with Yangzhou University and the Changsha University of Science and Technology. He has published over 300 international journals and conference papers. His research interests include wireless sensor networks, network performance analysis, and security. He is a member of the ACM.