Optimally-Weighted Multi-Scale Local Feature Fusion Network for Driver Distraction Recognition

LI SHAO FAN and GAO SHANGBING
Laboratory for Internet of Things and Mobile Internet Technology of Jiangsu Province, College of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an 223001, China
Corresponding author: Gao Shangbing (luxiaofen_2002@126.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1004904, in part by the Key University Science Research Project of Jiangsu Province under Grant 18KJA520001, and in part by the 333 Project Funding Program of Jiangsu Province under Grant BRA2018454.

ABSTRACT
Distracted driving is one of the main contributors to traffic accidents. In this work, we propose a novel multi-scale local feature fusion network for image-based distracted driver detection. Since the driver is the most important part to infer the distracted driver actions in a single image, our proposed method first detects the driver's body using person detection. Then capture abundant local body features after a repeated multi-scale feature fusion module. In addition to the features extracted from the whole image, our network also include the important feature of local body feature. The global feature and local feature are finally fused by an OAWS(optimally-weighted strategy). The experimental result shows that our methods achieve comparative performance on our own HY Large Vehicle Driver Dataset and the public AUC Driver Distracted Dataset.

INDEX TERMS
Convolution neural network, deep learning, distraction recognition, multi scale feature fusion.

I. INTRODUCTION
Distracted driving is the main cause of traffic accidents. According to data from the Department of Transportation, about 2 million traffic accidents occur every year, and over 80% of them are caused by distracted driving. In recent years, ADAS(Advanced Driver Assistance Systems) has been adopted by many vehicles, ADAS use a range of sensors to collect data inside and outside the vehicle, and also detect and recognize static and dynamic objects inside and outside the car. Driving behavior detection is the key technology in ADAS. This technique can effectively remind the driver to avoid traffic accidents. Therefore, driver distraction recognition has broad research prospects in computer vision and autonomous driving areas.

With the rapid development of deep learning and computer vision, many researchers have studied distracted driving detection in various ways. In recent years, deep learning has been widely used in the field of driver distracted detection. Compared with traditional algorithms, deep learning methods have been greatly improved in terms of performance and accuracy. In the application field of driver distracted recognition, various model has been proposed, such as AlexNet, VGGNet, GoogleNet, and ResNet, these have achieved excellent performance in computer vision task such as driver distracted recognition. But most of these methods use the global information of the image, it is easily affected by the noise from the complex driving environments. To extract more abundant and discriminate driver features and take into account the clues generated from the scene. As shown in Fig.1, we propose an optimally-weighted multi-scale local feature fusion network for driver distraction recognition. In this task, we extract the driver’s comprehensive multi-scale feature by person detection and utilize repeated feature fusion to get more low-level detail features and high-level abstract features. In addition, we proposed an optimally-weighted
module to focus on learning those representative global features and local features. Fig 1 is the overall framework of our proposed method. The major contribution of this article is summarized as follows:

1) A local multi-scale feature repeated fusion module is proposed to efficiently extract and aggregate the rich features from the driver’s low-level and high-level feature.

2) An optimally-weighted module to weigh the local body features and global features, so the model can fully consider the context information and learn more representative features.

3) The proposed method has trained end to end and achieve comparative performance on both our own HY Driver dataset and the public AUC Driver Distracted Dataset.

II. RELATED WORK

Due to the frequent occurrence of traffic accidents caused by distracted driving, distracted driving detection has attracted a lot of attention from industry and academic research groups. The following works are devoted to the detection of various distracted behaviors [1], [2], [3].

Using mobile phone is the main reason to increase the probability of traffic accidents. Seshadri et al. [4] created a dataset to detect the behavior of using mobile phones, and proposed detection method based on HOG. The designed AdaBoost classifier to get 93.9% accuracy. Le et al. [5] proposed a faster RCNN for hand and face detection, and achieved 94.2% accuracy, which is higher than all previous methods. Zhao et al. [6] proposed a hidden random forest method to detect drivers using small mobile phones.

Guo and Lai [7] used color and shape information to detect driving behavior. Yan et al. [8] combined a motion history map and pyramid gradient direction histogram for driver behavior recognition. Tran et al. [9] proposes a real-time distraction detection system using dual Cameras. The result shows that using two sources of data from different view cameras can achieve a better performance than using a single camera and achieve 96.7% accuracy in own dataset. Li et al. [10] design a lightweight network that combines depthwise convolution and pointwise convolution for detecting driver distraction in the TX2 target platform and achieves an accuracy of 95.98% accuracy when the latency is just about 32.8 ms. Baheti et al. [11] presents a modified VGG-16 network that uses various regularization techniques such as L2 regularization and achieves an improved accuracy of 96.13%. Koesdwiady et al. [12] used a fine-tune VGG-19 to detect Driver Distraction. It demonstrates that CNN achieves better performance than traditional algorithms in distracted recognition. Hu et al. [13] uses a multi-scale attention convolution neural network to refine the feature in image-based distracted recognition. Wagner et al. [14] proposes a driver action recognition system to detect cell phone usage and food consumption based on two IR cameras. The combined result achieves a test accuracy of 92.54% in its dataset. Baheti et al. [15] proposed a network structure called MobileVGG by replacing VGG’s all connect layers with a 1*1 convolutional layer and reduced to 512 neurons. This network achieves 95.24% accuracy on the AUC dataset.

Although the above research work has achieved good improvement results. However, most of the above works focus on using global information from the whole image. The core part of driver distracted recognition is the driver’s body, but the feature extraction of drivers is easily disturbed by noise in complex and changeable driving environments. To obtain more abundant features of the driver’s body, and also to consider the cues of the global background. Therefore, this paper proposed an optimally-weighted multi-scale local feature fusion network to recognize driver distracting action. While comprehensively extracting driver body features, the effect of global background on prediction is also taken into account.

III. THE PROPOSED METHOD

A. OVERVIEW

The method proposed in this paper is shown in Figure 2. Human body information plays an important role in this method. To locate the drive more accurately, this paper uses YOLOv5 to detect the driver’s human body and normalize the human body coordinates. The three-scale feature map of the backbone network is used to extract the driver features and send in repeated multi-scale feature fusion modules. Then perform weight strategy optimization fusion with the global features of the backbone network to identify the driver’s behavior. Our proposed method is mainly divided into two parallel branches, one branch utilizes the basic CNN to obtain the high-level feature expression map of the image. Another branch focuses on capturing rich multi-scale local driver’s scale features from low-level and high-level via YOLOv5 and multi-scale architecture. Finally, we also propose a weight optimization strategy that combines global average pooling and global max pooling.

B. GLOBAL FEATURE

The global feature plays an important role in driver behavior recognition, describing the relationship between driver actions and the whole image. We adopt the widely used CNN model ResNet50 as the backbone network. The whole image
L. S. Fan, G. Shangbing: Optimally-Weighted Multi-Scale Local Feature Fusion Network for Driver Distraction Recognition

FIGURE 2. Proposed optimally-weighted multi-scale local feature fusion network: (a) The proposed multi-scale local feature fusion module. (b) Detailed optimally-weighted strategy module.

is used as input to extract global features. The ResNet-50 used is pre-trained on ImageNet, the model is fine-tuned, and the last fully connected layer of the model is modified to adopt the driver behavior recognition task. After the entire image passes through the backbone network, the feature map \( F \) is obtained. Then followed by a residual block (Res) and get the global feature.

\[
f_C = \text{Res}_C(F) \tag{1}
\]

C. PERSON DETECTION

Since there are a large number of publicly available person image datasets and high-performance object detectors, it is easy to detect human bodies using existing object detectors. In this paper, YOLOv5 is used as the person body detector to predict the position of the human body in the image.

\[
\begin{align*}
b_x &= \sigma(t_x) + c_x \\
b_y &= \sigma(t_y) + c_y \\
b_w &= p_w e^{\omega} \\
b_h &= p_h e^{\beta}
\end{align*} \tag{2}
\]

where the YOLOv5 network take the whole image as input, and outputs a feature map consisting of cells, offset by \((c_x, c_y)\) which estimate \(k\) box coordinates \((t_x, t_y, t_w, t_h)\) prior width \(p_w\), \(p_h\). The detected human bounding box is \(x_0, b_x\) and \(b_y\) are the center points of the human bounding box, and \(b_w\) and \(b_h\) are the length and width of the human bounding box.

D. MULTI-SCALE LOCAL FEATURE FUSION MODULE

This branch focuses on learning the local abundant driver body feature. Following the person detection methods, we use region of interest (RoI) pooling on the driver regions to extract features. Then followed by a residual block (Res) and global average pooling (GAP) to extract a single-scale driver body feature \(f_h\).

\[
f_h = \text{GAP}(\text{Res}_h(\text{RoI}(F, x_0))) \tag{3}
\]

Multi-scale feature fusion can effectively aggregate the low-level detailed feature and high-level semantic information of images. The multi-scale feature fusion module needs three scale of input features from backbone network \(P_{\text{in}} = (P_{\text{in}l1}, P_{\text{in}l2}, P_{\text{in}l4})\), the output feature \(P_{\text{out}} = h(P_{\text{in}})\) and \(h()\) is the multi fusion operation. \(P_{\text{in}l0}\) represents a feature map with a resolution \(1/2^n\) of the input image. For example, if the input image resolution is \(512\times 512\), the \(P_{\text{in}l1}\) represent the feature map which resolution is \(256\times 256\) after downsampling operation. \(P_{\text{in}l2}\) represent the feature map which resolution is \(1/4\) of the original input image. The multi-scale repeated fusion operation combine different scale feature map.

\[
P_{\text{out}l1} = \text{Conv}(\text{UpSample}(P_{\text{in}l2} \oplus \text{UpSample}(P_{\text{in}l4} + P_{\text{in}l1})) \tag{4}
\]
FIGURE 3. Sample image of HY large vehicle driver dataset.

FIGURE 4. Sample image of AUC distracted driver dataset.

\[
P_{\text{out}}^{l_2} = \text{Conv}(\text{DownSample}(P_{\text{in}}^{l_1}) \oplus \text{UpSample}(P_{\text{in}}^{l_4}) + P_{\text{in}}^{l_2}) (5)
\]

\[
P_{\text{out}}^{l_4} = \text{Conv}(\text{DownSample}(P_{\text{in}}^{l_1}) \oplus \text{DownSample}(P_{\text{in}}^{l_2}) + P_{\text{in}}^{l_4}) (6)
\]

where the \text{Conv} represents convolution operation, \text{DownSample} represents downsampling operation, \text{UpSample} represents upsampling operation.

**E. OPTIMALLY-WEIGHTED MODULE**

Global features can provide some distinguishable cues for driver behavior recognition, while local human multi-scale fusion features provide rich human features. Normally, the two features are aggregated with GAP (global average pooling) or GMP (global max pooling) and used to replace the traditional fully connected layer. It can be regarded as a structured regularization that forces features to be mapped to the N classes. However, the use of GAP and GMP depends entirely on the specific task. GAP is an average aggregation of all values, which is more suitable for local human features, but there is also a problem that GAP always pays too much attention to those frequently occurring patches, whereas GMP is the opposite. GMP only pays attention to the largest value in the feature, but it also loses some important information for action recognition. The two are a complementary relationship. Therefore, in order to obtain more comprehensive and reasonable information, this paper introduces weights matrix \( \rho \) to the global feature features and allocates them reasonably.

\[
\hat{f} = f_{\text{fusion}}^h + f_C
\]

\[
\rho = \text{Softmax}(W_{\omega\rho} \hat{f} + b_{\omega})
\]

\[
\text{score} = \text{Softmax}(\rho \hat{f}_{\text{GMP}} + (1 - \rho) \hat{f}_{\text{GAP}})
\]

where the weight matrix \( W_{\omega\rho} \) and bias \( b_{\omega} \) are learnable parameters. \( \hat{f}_{\text{GMP}} \) is the feature map after global max pooling and the \( \hat{f}_{\text{GAP}} \) is the feature map after global average pooling.

**IV. EXPERIMENTS**

**A. DATASET**

1) HY LARGE VEHICLE DRIVER DATASET

The dataset comes from the in-vehicle surveillance video of trucks and buses provided by our cooperative companies in the industry. The cameras are installed in different positions in the vehicle, and the environment inside the vehicle
is complex and changeable, which effectively improves the richness of the data. Figure 3 shows some sample images of the five actions in the dataset. Firstly, the long surveillance video is cut into short videos with distracting behaviors, and then the short videos are divided into video frames. The representative frames in video frames are selected as the dataset. The human body is annotated by combining YOLOv5 automatic detection and manual manual annotation. The dataset contains 43776 images in total, 38756 are used as training sets and 5020 are used as the test set. Table 1 shows the details of the dataset.

2) AUC DISTRACTED DRIVER DATASET
AUC-V1 dataset consists of 31 participants from 7 countries. It includes a total of 12977 training images and 4331 testing images. The dataset is divided into 10 categories. The categories are (1) C0: safe driving; C1: Text right; C2: Talk right; C3: Text left; C4: Talk left; C5: Adjust radio; C6: Drink; C7: reaching behind; C8: Hair and makeup; C9: Tall to passenger. AUC-V2 dataset is an enhanced version of the previous dataset, and the data includes a total of 44 people from 7 countries. Among them, 12555 images are used for training and 1923 images are used for testing. Some sample data are shown in Fig. 4.

B. IMPLEMENTATION DETAILS
For global feature, we use ResNet-50 as a backbone network. The backbone CNNs are initialized with the pre-trained ImageNet weights. The input size of image is scaled to 224*224. For local features, YOLOv5 is used to detect the position of the human body frame, and the position of the human body is normalized to crop local features at different scales of the backbone network. The method of local feature extraction is ROI+ Pooling operation, and the backbone network extracts the features of the three scales are 28*28, 14*14, and 7*7, respectively. After repeated multi-scale feature fusion, the features of the three scales still maintain the original resolution.

The network training uses the cross-entropy loss function to update the weight of the network model. The initial learning rate is set to 1e-3, and the batch size of the algorithm training is 64. The Stochastic Gradient Descent optimizer with a momentum of 0.99 is applied to the cross-entropy loss function. For the SGD, a multi-step learning rate with parameters and calculations, which is not conducive to the practicability of the algorithm. While [17] and [19] use sliding window detectors to detect human body parts, this object detection algorithm is easily affected by noise in complex environments, and will seriously affect the recognition accuracy of next-stage behaviors.

The result shows that our method can perform better than both DenseNet and AlexNet. Although our method obtained 95.84% which is close to 96.31% of the best method on Table 1 shows the details of the dataset.

### Table 1. Result of HY large vehicle driver dataset.

| Method       | C0 | C1 | C2  | C3  | C4  | accuracy |
|--------------|----|----|-----|-----|-----|----------|
| AlexNet      | 58.51 | 73.73 | 95.37 | 83.72 | 84.74 | 79.06    |
| DenseNet     | 84.23 | 73.61 | 94.93 | 96.37 | 0.8298 | 87.52    |
| ResNet-50    | 82.25 | 86.47 | 99.49 | 95.93 | 82.35 | 89.52    |
| Fusion[11]   | 90.95 | 72.51 | 91.15 | 97.47 | 95.23 | 90.53    |
| Ours         | 95.69 | 86.70 | 94.97 | 97.58 | 93.40 | 94.34    |

for precision and recall are:

\[
\text{Precision} = \frac{TP + TN}{TP + FN + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]
TABLE 3. Result of AUC distracted driver dataset.

| Dataset  | Method                   | Accuracy |
|----------|--------------------------|----------|
| AUCV1    | 1D-two stream[16]        | 77.10    |
|          | Fusion[21]               | 92.36    |
|          | AlexNet+HOG features[17] | 93.39    |
|          | DenseNet+Latent Pose[18] | 94.20    |
|          | VGG-16 with Regularization[11] | 96.31 |
|          | GA-Weighted Ensemble[19] | 95.31    |
|          | Ours                     | 95.84    |
| AUCV2    | Inception-V3[20]         | 90.07    |
|          | ours                     | 90.25    |

AUCV1. but the accuracy of our method achieve the best accuracy in AUCV2.

Fig5, Fig6 and Fig 7 are the confusion matrices tested by this method on HY Large Vehicle Driver Dataset and AUC datasets, where the left row represents the true class, and the column represents the predicted class. The values in each square of the matrix represent the proportion of the predicted category and the corresponding number of samples, respectively.

In the confusion matrix, the values on the diagonal are the correct predictions, and the other values are the wrong predictions. On the data set of our Large Vehicle Driver dataset, the calculation shows that the overall accuracy and recall rate of this method is both 94.30%, but the recall rate of looking around is only 86.7%, of which 9.98% of the 90 samples were incorrectly predicted as safe driving, this is because although this method obtains the feature of the driver through the prior method of person detection, and eliminates the interference problem caused by the noise of backgrounds, however, the changes in the posture of the driver looking around and safe driving is very small, and the main difference is the change in the driver’s face, so the model is easily confused between the predictions of these two categories. The recall rate of talking on the phone is 94.97%, of which 4.74% of the 129 images in the sample are mistakenly classified as smoking, while the recall rate of smoking is 93.4% and 5.21% of the 83 images are confused as calling. This is because the postures of smoking and making a phone call are very similar, but both cigarette and mobile phone are relatively small objects, which are easily ignored in the convolution process of the neural network, so it is difficult for the model to distinguish the difference between the two categories.

On the AUCV1 public data set, the overall accuracy and recall rate of this method is both 95.61%. The recall rate of reaching behind and makeup is relatively low, which are 93.69% and 94.14%, respectively, and 15 and 11 images are wrongly predicted as safe driving. This is because some images of this category in the test set have relatively small pose changes, so the model will produce some misjudgments, but AUC overall, because the test set is not very large, the
overall test performance of the model is better. Only a small number of pictures are mispredicted. Although high accuracy was achieved on both datasets, but how to make the neural network model pay more fine-grained attention to the small intra-class gaps in these parts under specific tasks is an issue to be focused on in the future.

D. ABSTRACTION STUDY
We have also conducted several experiments to analyze the effectiveness of the components on our method on three datasets (HY Large Vehicle, AUC-V1, and AUC-V2). Where G represents the global feature, LF represents the local fusion module, and OW represents the optimally-weighted module. We evaluate the impact of the components on our method. The result shown in Table 4, the performance of local fusion features is significantly improved compared to global features in HY Large Vehicle Driver Dataset. However, the improvement in accuracy on the AUC dataset is not noticeable. Since the viewpoint and environment in HY Large Vehicle dataset is more complex than the AUC dataset. So person detection reduces the noise from the complex driving environment and changeable viewpoint. Since the local feature fusion can get more abundant body features from low level and high level. When combination of three components, the model can achieve best accuracy. The results shows that the driver distracted recognition combined with human detection has better performance when facing complex driving environments and different viewpoint.

The results of the ablation study show the effectiveness of the proposed component, which can improve recognition accuracy. While combining the local fusion module and optimally-weighted module, the accuracy is 4.8% higher than the global feature in HY Large Vehicle Driver Dataset. More than 3% improvements have been achieved in the AUC dataset.

V. CONCLUSION
This paper proposes optimally-weighted multi-scale local feature fusion network. While combining person detection and local multi-scale repeated feature fusion structure to obtain rich human features, the weight optimization strategy with GMP and GAP is also used to focus on learning those representative global features and local human features. While fully considering the human-centered driving behavior recognition, this method also pays attention to the global context clues, and achieves comparative results on both the self-built HY Large Vehicle Driver Dataset and the public AUC Driver Distracted Dataset. However, how to use the driver’s body keypoint position to obtain more detailed and discriminative visual and spatial features is our next research work.

REFERENCES
[1] Y. Xing, J. Tang, H. Liu, C. Lv, D. Cao, E. Velenis, and F. Y. Wang, “End-to-end driving activities and secondary tasks recognition using deep convolutional neural network and transfer learning,” in Proc. 23th Int. Intell. Vehicles. Symp. (IV), Chongshina, China, Jun. 2018, pp. 1626–1631.
[2] Y. Liao, S. E. Li, W. Wang, Y. Wang, G. Li, and B. Cheng, “Detection of driver cognitive distraction: A comparison study of stop-controlled intersection and speed-limited highway,” IEEE Trans. Intell. Transp. Syst., vol. 17, no. 6, pp. 1628–1637, Jun. 2016.
[3] S. Park, F. Pan, S. Kang, and C. D. Yoo, “Driver drowsiness detection system based on feature representation learning using various deep networks,” in Proc. Asian Conf. Comput. Vis., Macao, China, Nov. 2016, pp. 154–164.
[4] K. Seshadri, F. Juefei-Xu, D. K. Pal, M. Savvides, and C. P. Thor, “Driver cell phone usage detection on strategic highway research program (SHRP2) face view videos,” in Proc. IEEE Comput. Vis. Pattern Recognit. Workshops, Boston, MA, USA, Jun. 2015, pp. 35–43.
[5] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides, “Multiple scale faster-RCNN approach to driver’s cell-phone usage and hands on steering wheel detection,” in Proc. IEEE Conf. (CVPRW), Las Vegas, NV, USA, Jun. 2016, pp. 46–53.
[6] C. H. Zhao, B. L. Zhang, J. He, and J. Lian, “Recognition of driving postures by contourlet transform and random forests,” IET Intell. Transp. Syst., vol. 6, no. 2, pp. 161–168, 2012.
[7] G. Guo and A. Lai, “A survey on still image based human action recognition,” Pattern Recognit., vol. 47, no. 10, pp. 3343–3361, Oct. 2014.
[8] C. Yan, F. Coenen, and B. Zhang, “Driving posture recognition by joint application of motion history image and pyramid histogram of oriented gradients,” Int. J. Veh. Technol., vol. 2014, pp. 1–11, Jan. 2014.
[9] D. Tran, H. M. Do, J. Lu, and W. Sheng, “Real-time detection of distracted driving using dual cameras,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Las Vegas, NV, USA, Oct. 2020, pp. 2014–2019.
[10] P. Li, Y. Yang, R. Grouu, G. Wang, R. Li, Y. Wu, and Z. Huang, “Driver distraction detection using octave-like convolutional neural network,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 7, pp. 8823–8833, Jul. 2022.
[11] B. Baheti, S. Gajre, and S. Talbar, “Detection of distracted driver using convolutional neural network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Salt Lake City, UT, USA, Jun. 2018, pp. 1032–1038.
[12] A. Koeswidiwaty, S. M. Bedawi, C. Ou, and F. Karray, “End-to-end deep learning for driver distraction recognition,” in Proc. 14th Int. IEEE Conf. Image Anal. Recognit. (ICIAR), Montreal, QC, Canada, Jul. 2017, pp. 11–18.
[13] Y. Hu, M. Lu, and X. Lu, “Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network,” Signal Process., Image Commun., vol. 81, Feb. 2020, Art. no. 115697.
[14] B. Wagner, F. Taffner, S. Karaca, and L. Karge, “Vision based detection of driver cell phone usage and food consumption,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 5, pp. 4257–4266, May 2022.
[15] B. Baheti, S. Talbar, and S. Gajre, “Towards computationally efficient and realtime distracted driver detection with MobileVGG network,” IEEE Trans. Intell. Vehicles, vol. 5, no. 4, pp. 565–574, Dec. 2020.

TABLE 4. Ablation study on three different datasets.

| Dataset   | Component | Accuracy |
|-----------|-----------|----------|
| HY Driver | G, LF, OW | 90.34    |
| AUC-V1    | G, LF, OW | 95.84    |
| AUC-V2    | G, LF, OW | 92.42    |
[16] N. Moslemi, R. Azmi, and M. Soryani, “Driver distraction recognition using 3D convolutional neural networks,” in Proc. 4th Int. Conf. Pattern Recognit. Image Anal. (IPRIA), Mar. 2019, pp. 145–151.

[17] M. R. Arefin, F. Makhmudkhujaev, O. Chae, and J. Kim, “Aggregating CNN and HOG features for real-time distracted driver detection,” in Proc. IEEE Int. Conf. Consum. Electron. (ICCE), Las Vegas, NV, USA, Jan. 2019, pp. 1–3.

[18] A. Behera and A. H. Keidel, “Latent body-pose guided DenseNet for recognizing driver’s fine-grained secondary activities,” in Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Auckland, New Zealand, Nov. 2018, pp. 1–6.

[19] Y. Abouelnaga, H. M. Eraki, and M. N. Moustafa, “Real-time distracted driver posture classification,” in Proc. 32th Int. Neural. Inf. Syst. (NIPS), Montreal, QC, Canada, Dec. 2018, pp. 1–8.

[20] H. M. Eraki, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, “Driver distraction identification with an ensemble of convolutional neural networks,” J. Adv. Transp., vol. 2019, pp. 1–12, Feb. 2019.

[21] M. Alotaibi and B. Alotaibi, “Distracted driver classification using deep learning,” Signal, Image Video Process., vol. 14, no. 3, pp. 617–624, Apr. 2020.

LI SHAO FAN received the B.S. degree from the Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai’an, China, in 2020, where he is currently pursuing the master’s degree with the Faculty of Computer and Software Engineering. His recent research interests include computer vision and deep learning.

GAO SHANGBING received the Ph.D. degree in computer science and technology from the Nanjing University of Science and Technology, China, in 2014. He is currently a Professor with the Faculty of Computer and Software, Huaiyin Institute of Technology, China. He also holds a Postdoctoral Researcher with the Faculty of Computer Science and Engineering, South East University. His current research interests include pattern recognition and computer vision.

***

VOLUME 10, 2022