Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

A Lightweight Architecture For Driver Status Monitoring Via Convolutional Neural Networks

Jinyang He[1], Jiqing Chen[2], Jun Liu[1,3], and Hengyu Li[1]

***Abstract*—Nowadays, a significant portion of road accidents is attributed to dirver's inattention or dangerous behavior such as call or smoke. Monitoring the driver's status through computer vision techniques to detect driving errors in real time to avoid vehicle collisions is one of the most important issues in machine vision and Advanced Driver Assistant System(ADAS). In this paper, we propose a novel facial landmark model and a lightweight architecture of Driver Status Monitoring(DSM) by integrating several deep learning models in order to monitor the driver's inattention caused by drowsiness, fatigue and distraction in real time. With inverted residual block and multi-scale feature map, the proposed facial landmark model achieve 4.90% normalized mean error on common set of 300W dataset. The proposed detection model achieved a mean average precise(mAP@0.5) of 90% on our validation dataset. Several results presented in this paper may provide meaningful insights for the autonomous driving research community and automotive industry for future algorithm development.**

***Index Terms*—Driver status monitoring, Deep Learning, Advanced Driver Assistant System**

I. Introduction

Nowdays, autopilot technology creates a safer driving environment by providing the ability to learn from driving experience for autonomous vehicles and avoiding human error. The car manufacturers are also working hard to make cars completely safe. With the development of Advanced Driver Assistant System(ADAS), vehicles can monitor driver performance, alertness and driving intentions to improve traffic safety through a so-called driver state monitoring(DSM) system.

Driver's status is crucial because one of the leading causes of traffic accidents is relevant to driver's distraction or fatigue. According to the statistics of the World Health Organization(WHO), more than 1.2 million people ard killed and up to 50 million people are injured due to road accidents [1]. For a long time, fatigue or distraction detection has always been a major concern. However, some dangerous behaviors such as call or smoke have become a vital obstacle to traffic safety. For these reasons, driver status monitoring(DSM) system which can detect unsafe actions of the driver is highly requisite. And a number of crashes can be avoided by analyzing driver status and alerting them through dedicated sensors. These techniques provide real-time performance to allow the driver to immediately control the vehicle in an emergency and ensure safe driving [2].

Fig. 1. The instrumented vehicle with vision system [6].

Over the past few decades, diverse techniques have been applied to measure dirver's status such as fatigue or distraction. Some of the methods focus on the driving pattern of the vehicle. These methods focus on the performance of steering-wheel control or the increasing times of lane deviation [3]. And there are some techniques exploit bioelectrical signal such as ECG, EEG, EOG [4]. However, the techniques mentioned above have significant limitations. Either not robust enough or too complicated for practial purposes. Another popular means of implementing DSM is based on computer vision. Several methods are proposed and researched in the literature to detect driver's distraction by means of gaze patterns and head movements [5]. Drowsiness and yawn always serve as fatigue indicators. The frequency of eye-blinking which measures the percentage of time when the eye is more than 80% closed have also been an important indicator of fatigue. The computer vision technology focuses mainly on eye closure, yawning patterns, facial expressions and head movement detection.

In this work, we propose a lightweight facial landmark model which consist of several inverted residual blocks. The inverted residual block reduce the loss of information of the feature vector through the activation function, and it is light enough by means of using depthwise convolution. Unlike other regression-based methods that use single-scale feature map for prediction, multi-scale feature map is used for more accurate

The corresponding author is Hengyu Li, email: lihengyu@shu.edu.cn

This research was funded by the National Natural Science Foundation of China (Grant Numbers 61525305, 61703181 and 61625304), the Shanghai Natural Science Foundation (Grant Numbers 17ZR1409700 and 18ZR1415300), the basic research project of Shanghai Municipal Science and Technology Commission (Grant Number 16JC1400900).

[1]School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[2]School of Mechanical Engineering, Guangxi University, Guangxi, China
[3]Department of Mathematics, Jining University, Shandong, China

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

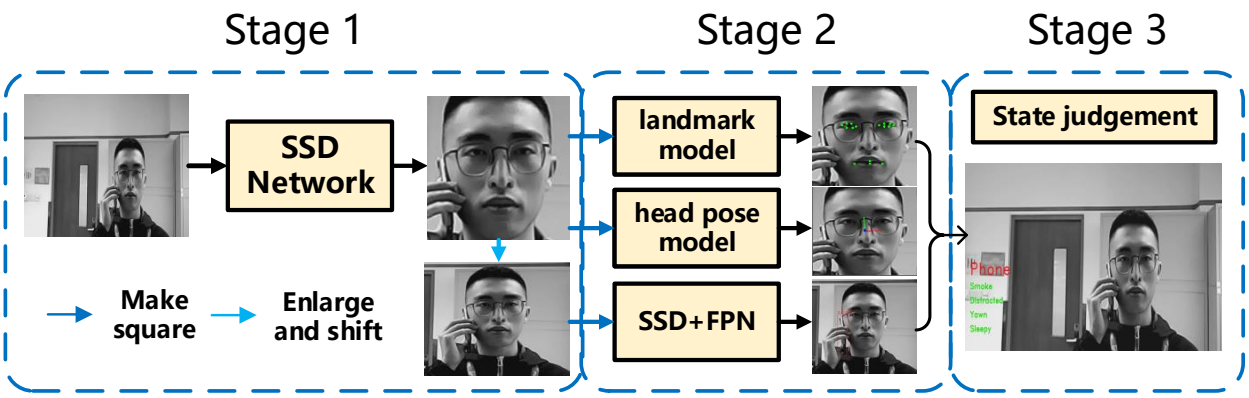

Fig. 2. The overall architecture of DSM: Stage 1 consists of face detector and preprocess of face bounding box. The facial attributes such as landmarks and head pose are extracted in the Stage 2. The phone-smoke detection also implemented in the second stage. Stage 3 take all the information from Stage 2 as input and make state judgement, then give corresponding warnings.

results in our method. Benefit from these means, our landmark model can achieve a better accuracy-speed tradeoff. For detection, although the performance of the SSD [7] detector is not as well as that of the general Faster R-CNN [8] or R-FCN [9], but show competitive capability with lightweight backbone. So we choose SSD detector and lightweight backbone MobileNetV2 [10] as our baseline. While the shallow features with small receptive field have poor feature representative ability so that the small size of smoke and varied appearance of phone are key points to be solved. For this problem, the Feature Pyrimaid Network [11] is applied to improve SSD's poor performance for small object and get comparable results.

Meanwhile, we proposed a lightweight architecture of DSM by integrating several proposed model and some state-of-the-art algorithms in deep learning and computer vision. All deep learning based models are optimizied by the OpenVINO Toolkit [12] so that the whole workload can be highly accelerated and effective on hardware. The entire pipeline can run in real time on the cpu or embedded board while preserving reasonable accuracy. Our method take the real-time video frames captured by a single infrared camera as input and detect five status(distraction, drowsiness, yawn, listening call, smoking). The Figure 1 show the vision system and the installation position of the camera in the car.

This paper is organized as follows. In Section II, we presents the architecture of DSM and several proposed models which achieve a speed/accuracy trade-off. And we describe some implementing details including model training and inference acceleration in Section III. Finally, this paper is summarized in Section IV where we discuss some technological difficulties and possible work to be done in the future.

II. Methodology

For safe driving, it is essential to monitor the status of driver especially on special occasions such as freight transport. This section describes the proposed architecture of DSM first.

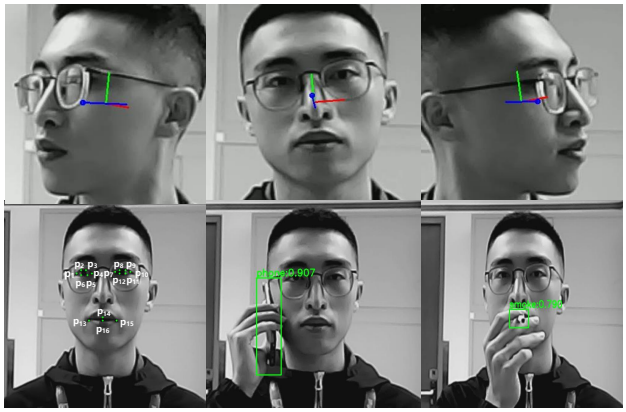

Fig. 3. Several detection results of our deep learning models. The top half of picture shows the results of head pose. The following is 16 facial landmarks including eyes and mouth. And two dangerous behaviors such as listening to phone and smoking.

Afterwards, the phone-smoke detection model is introduced. The entire architecture contains four deep learning networks, which are face detection, facial landmark detection, head pose detection and phone-smoke detection network. All the models are optimizied by the OpenVINO Toolkit. The whole workload can run in real time on the CPU.

A. Overall Architecture

The overall architecture of DSM consists of three stages as illustrated in Figure 2. Firstly, the face detection model take the frames captured by near-infrared camera as input and find the faces. We chose a popular SSD-MobileNet detection model from the Open Model Zoo [13] repository to detect faces [14]. Considering that driver's face may not the only face in the picture, we will save the largest face bounding box found for the subsequent model. In order to improve the performance of following model, we do some preprocessing on the input

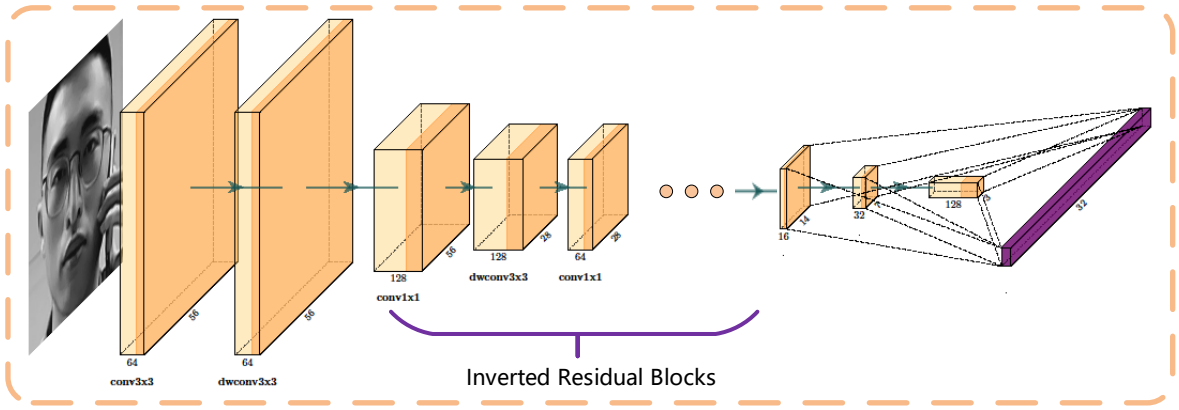

Fig. 4. Overall structure of the facial landmark model. The input resolution of model is 112 x 112. Similar to MobileNetV2, there are several inverted residual blocks for feature extraction. Finally, multi-scale feature map is used as the input to fully-connected layer for regressing landmarks.

face bounding box of these models. The immediate purpose of preprocessing is to eliminate some background interference to get a more discriminative region, usually by cutting or changing the shape of the image with certain rules. For landmark and head pose model, we make the facial bounding box to square and side length is the maximum of length and width of bounding box. And the original face bounding box is enlarged by 1.8 times and shifted the center of the box along the positive direction of the y-axis for input of phone-smoke dectection model. This method can highly improve the performance and efficiency.

In the second stage, the subsequent networks take the face bouning box after preprocessing as input and output corresponding information including facial landmark, head pose, and detection results. Specifically, the facial landmark detection network take the square face bounding box after resizing and output 16 coordinates of landmarks. Head pose is represented by the euler angles(yaw, pitch or roll) which are estimated by the network as shown in the first line of Figure 3. The supported degree range is $[-90, 90]$ for yaw and $[-70, 70]$ for pitch and roll. The phone-smoke detection network are used to detect the calls and smoking. It take the enlarged bounding box as search space which can highly reduce the redundancy computations compared to use whole image during training and inference phase. And there are several detection results are presented in Figure 3.

The last one is state judgement stage which can monitor the real-time status of driver through the information extracted from the second phase. A driver's status are determined by various indicators in different work. In this study, undersize eye aspect radio(EAR) which continuously occurr in multiple frames serve as indicator of drowsiness. Similarly, the oversize mouth aspect radio(MAR) is used to detect yawn. Both EAR and MAR can obtain from the facial landmarks. The EAR and MAR is defined as

$$EAR = \frac{\|p_2 - p_6\|_2 + \|p_3 - p_5\|_2}{2\|p_1 - p_4\|_2}$$

and

$$MAR = \frac{\|p_{14} - p_{16}\|_2}{\|p_{13} - p_{15}\|_2}$$

where p_i is the corresponding landmark in Figure 3. Distraction is detected by the head pose which is obtained from head pose detection model. The dangerous behaviors such as listening to phone or smoking are also warned by the monitoring system. Overall, the lightweight architecture allow DSM system to monitor the driver's abnormal status and dangerous behavior in real time, and give appropriate warnings in a timely manner.

B. Facial landmarks model

As other deep learning based models, several convolutional layers be used to extract features and predict landmarks in a regression manner. Inspired by the MobileNetV2, we use several inverted residual blocks as our backbone network. By using lightweight backbone network, the computational complexity is significantly reduced and the speed is increased. Moreover, we can reduce the width parameters of the inverted residual blocks to further accelerate and compress the network without obvious accuracy degradation. In order to get better results, we use the multi-scale feature map as the intput of the final fully-connected layer. The output of fully-connected layer is 16 coordinates of landmark $[x_1, y_1 \cdots x_{16}, y_{16}]$. The loss function is defined as

$$L(P^*, P^{GT}) = \frac{1}{M \times N} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\|\hat{p_{m,n}} - \vec{p_{m,n}}\|}{d_m}$$

or

$$L(P^*, P^{GT}) = \frac{1}{M \times N} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\|\hat{p_{m,n}} - \vec{p_{m,n}}\|_2^2}{d_m^2}$$

where $P^* = \{\vec{p}^*_{m,n}\}_{n=1:N}$ is the predicted landmarks of network. The d_m which denotes outer-eye-corner distance of m-th sample are used for normalizing. In Figure 4, we show the structure of the facial landmark model.

Fig. 5. Custom dataset for phone or smoke detection.

C. Phone-smoke detection model

Speed/accuracy trade-offs is vital for model selection. There are many popular detectors such as Faster R-CNN, R-FCN, YOLO, and SSD. In [5], a number of comparisons are made between different detectors. The results show that SSD performs not as well as two-stage detectors like Faster R-CNN or R-FCN in general, but outperforms them with lightweight backbone. According to this fact, the lightweight backbone MobileNetV2 are used inside SSD detector since it has low computational complexity and less parameters. Therefore we chose SSD-MobileNetV2 as our baseline model.

However, the performance of general SSD model is unsatisfactory for small objects such as smoke in our task. Since the shallow feature maps which have small receptive field are poor feature representation capacity. Therefore, we introduce the Feature Pyrimaid Network(FPN) into the SSD-MobileNetV2 to enhance the feature representation of shallow features and achieve better accuracy for smoke dectection. We denote the output of these last inverted residual blocks as $\{C_3, C_4, C_5, C_6, C_7\}$, and note that they have stride of $8, 16, 32, 64, 128$. As shown in Figure 6, the upsampled map is merged with the corresponding bottom-up feature map by element-wise addition. Finally, the set of feature maps is called $\{P_3, P_4, P_5, P_6, P_7\}$ for prediction. We introduce the focal loss [15] deriving from the cross entropy (CE) loss for binary classification:

$$CE(p, y) = \begin{cases} -log(p) & if\ y = 1 \\ -log(1-p) & otherwise \end{cases}$$

where $y \in \{\pm 1\}$ groundtruth class and $p \in [0, 1]$ is the estimated probability for the groundtruth class with label $y = 1$. For convenience of expression p_t is defined as:

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases}$$

and rewrite $CE(p, y) = CE(p_t) = -log(p_t)$.

We use an α-balanced variant of the focal loss:

$$FL(p_t) = -\alpha_t (1-p_t)^{\gamma} log(p_t)$$

In the above focusing parameter γ smoothly adjusts the rate at which easy examples are downweighted. The weighting factor $\alpha_t \in [0, 1]$ is introduced to address class imbalance. The focal loss solve the problem of positive and negative sample imbalance and accelerate convergence.

III. Implementation Details

In this section, implementation details and experimental results are presented. Dataset and training process are described in the first place.Secondly, the use of hardware and software are explained in details. To evaluate the performance of our proposed models, we show several results in the end.

A. Dataset and Training

All faces are cropped according to a given bounding box and resized to 112 x 112 for pre-processing during the training of facial landmark model. A variety of facial poses are usually one of the challenges for facial landmarks detection. We show some popular dataset statistics in Table 1. Since excessive head pose will be detected by the head pose model, the range $[-45°, 45°]$ of pose is more in line with our application scenario. The 300W dataset including AFW, LFPW, HELEN, XM2VTS and IBUG with 68 landmarks is used for training and validation. We utilize 3,148 images for training and 689 images for testing. The testing images are divided into two subsets, say the common subset formed by 554 images from LFPW and HELEN, and the challenging subset by 135 images from IBUG. The common and the challenging subsets form the full testing set [16].

TABLE I
SUMMARY OF THE MOST POPULAR FACE LANDMARK DATASETS.

Dataset	Size	pose	points
300W	3,837	$[-45°, 45°]$	68
300-VW	218,595	$[-45°, 45°]$	68
AFLW	24,386	$[-90°, 90°]$	21

For facial landmark model, we increased the training data by flipping and rotating them from $-30°$ to $30°$ with an interval of $5°$. All the images are resized to 112x112 and converted to grayscale with three channels in order to be consistent with our data captured by infrared camera. No pre-training is done to the proposed models and the training is done completely from scratch. Starting learning rate of 10^{-4} is taken since the initial weights are all randomly initialized.We employ the optimizer Adam for opimization with the weight decay of 10^{-6}. The

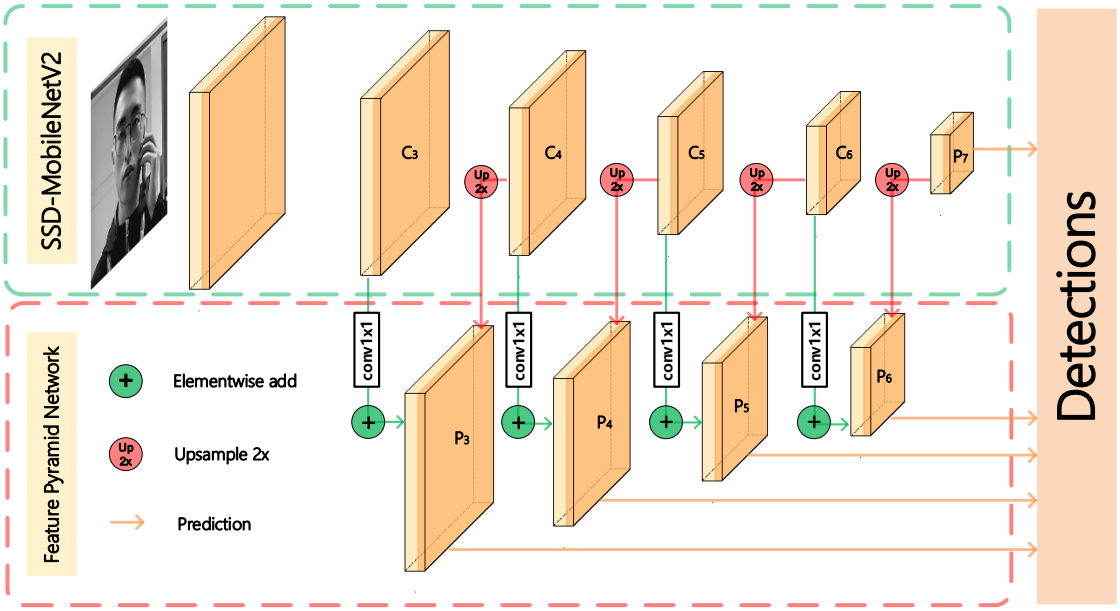

Fig. 6. SSD-MobileNetV2 with Feature Pyrimaid Network.

maximum number of iterations is 64K, batch size is 256 for faster convergence and better results.

The experiment on phone-smoke model was performed on the custom dataset. The training data of phone-smoke detection model are collected from various scenes and individuals with infrared camera. We collect multiple video clips captured by the WIN10 camera software. All the pictures were recorded in 1280x720 resolutions. There are two states of behaviors including call and smoke in each video. Finally, we select 1451 individual frames from all the videos to form our initial dataset. The custom dataset contains 775 call images and 676 smoking images from more than 30 people. Since custom dataset is from videos, we cannot randomly divide all the pictures into training set and validation set. We randomly selected a total of 114 pictures from 5 people as a validation set, and the rest 1337 images as training set. Figure 5 shows example images of our data.

For phone-smoke detection model, the offline and online data augmentation are tried to increase data since the limited amount of training data. In offline data augmentation, all the images are resized to 300x300 resolution which is the input size of model. Then we attempted a sliding-window crop strategy which ensure at least one complete object in the cropped image. In our experiment, the online augmentation with random flip and random crop is more effective than offline augmentation. In the experiment, TensorFlow is used for training and evaluation. Our model uses SSD as detector and MobileNetV2 as the backbone network. The focusing parameter γ is set to 2 and α_t is set to 0.75. The model is fine-tuned on weights which pre-trained on COCO dataset. We use the optimizer RMSPROP and initial learning rate is 0.04 with 5000 steps warmup. Then we train the model with batchsize of 24 and 150K iterations on two GPUs.

B. Hardware and Software Environments

The GTX 2080Ti GPU was used for training and evaluation. The GTX 2080Ti has 4352 cuda cores with a base frequency 1350MHz and a boost frequency 1545MHz. The PC has Intel i7-7800X CPU with 3.5GHz and 32GB of RAM. The near-infrared sensor for image capturing is 720p high-definition camera from RMONCAM corporation. Ubuntu 16.04 for OS and TensorFlow for deep learning framework were used. Further we use the Tensorflow Object Detection API [17]to construct training and testing pipeline simply. The version of OpenVINO Toolkit is latest 2019.R1.1. The whole dataset is convert to TFRecord format for fast and effective data processing. C++ programming language, OpenCV(Open Source Computer Vision Library), and OpenVINO Toolkit have been used for a complete implementation of our software system.

C. Experiment Results

Firstly, we evaluate the performance of the proposed facial landmark model. The evaluation metrics is normalized mean error which averages normalized errors over all groundtruth landmarks. There are two common normalizing factors which are eye-center-distance as the inter-pupil normalizing factor and the outer-eye-corner distance denoted as inter-ocular. For 300W, we adopts the inter-ocular factor and report corresponding results on the test dataset. Moreover, a single prediction result is treated as failure if accumulation of errors of all landmarks is larger than 0.1. The failure rate is also used as a metric in our experiment and the results have shown in Figure 7.

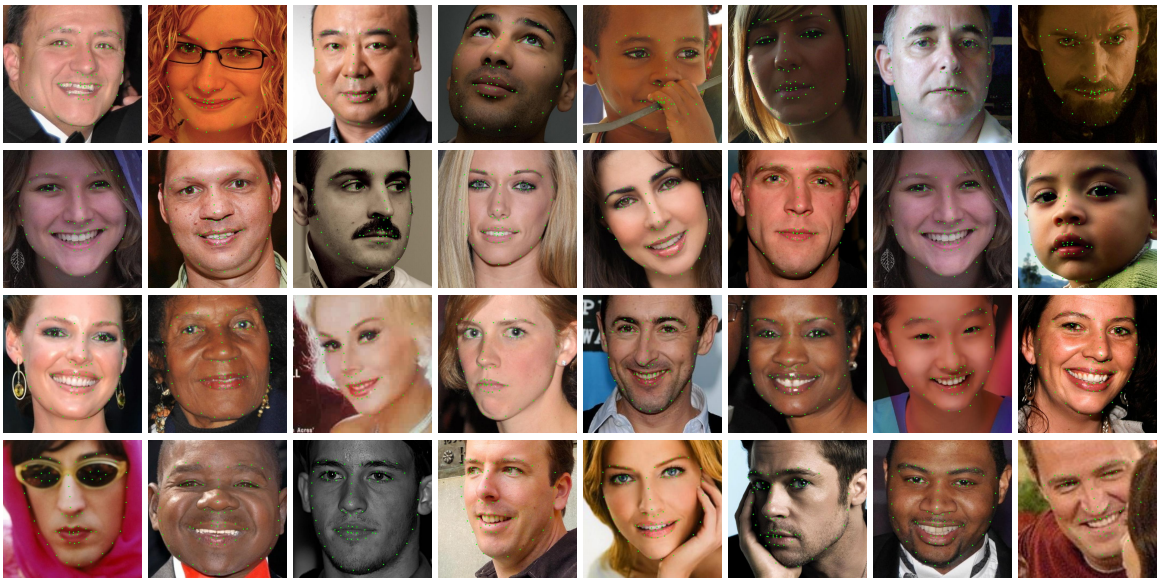

Fig. 7. Several experiment results on 300W dataset by our LandmarkNet-multi.

To further demonstrate the performance of the proposed landmark model, we show some comparative experiments on the 300W dataset with 68 landmarks. The Ablation study about multi-scale(LandmarkNet-multi) versus single-scale(LandmarkNet-single) feature map is also presented in Table 2.

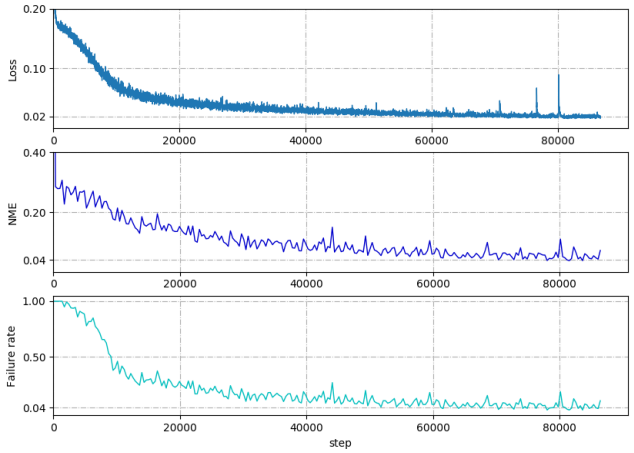

Fig. 8. Train and evaluation statistics. (a) Loss curve (b) Normalized mean error(Inter-ocular Normalization). (c) Failure rate(10%).

For phone-smoke detection model, COCO detection metric mAP@0.5 is adopted for evaluation the performance. A prediction will be considered as a true positive sample if the intersection over union(IOU) with groundtruth is more than 0.5 in the experiment. And we evaluate our model on the custom validation dataset which include 114 pictures from 5 people.There are several results including loss curve and mAP performance have been shown in Figure 8.

TABLE II
COMPARISON ABOUT NORMALIZED MEAN ERROR ON THE 300W DATASET.

Method \ Subset	Common	Challenge	Full
Normalized Mean Error(NME)			
PIFA-CNN [18]	5.43	9.88	6.30
RDR [19]	5.03	8.95	5.80
LandmarkNet-single	6.04	12.01	7.22
LandmarkNet-multi	**4.90**	**9.08**	**5.73**
LandmarkNet-16	3.75	—	—

IV. CONCLUSIONS

This paper propose a novel architecture of DSM which detect driver's error or dangerous behaviors. We train and integrate several lightweight deep learning models to compose the monitoring system. The whole architecture only need a single near-infrared camera for capturing video frames while can run on CPU or mobile devices in real time. The face attribute information is obtained by face detection model. Meanwhile, we detect dangerous behaviors such as smoking and making calls from enlarged areas around the face. The final stage gives the driver a corresponding prompt and warning based on all the acquired status information.

Several experiment results on custom dataset show the effectiveness of phone-smoke model as well as the architecture. And We use OpenVINO Toolkit to further accelerate the inference speed so that the entire architecture can run in real time on a CPU or embedded device. The optimized deep learning model can achieve a great accuracy-speed tradeoff when used in driver status monitoring system.

For future work, we will improve the detectability of the face attribute model, especially under some harsh conditions

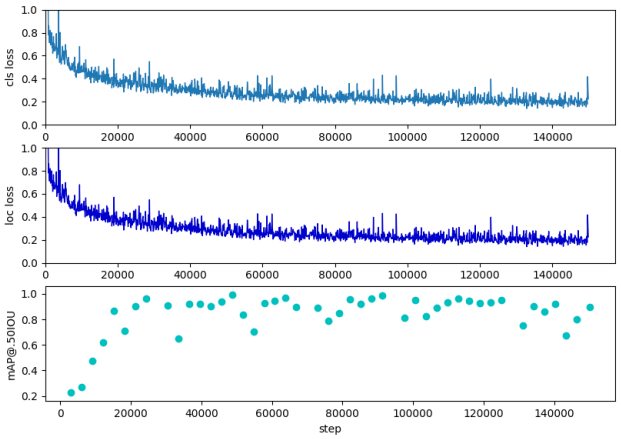

Fig. 9. Loss and mAP curve. (a) Classfication loss (b) Localization loss. (c) mAP@0.5IOU.

such as low light or different expressions. Moreover, the time information should be used by deep learning models to obtain better robust monitoring results in videos. So action recognition is one of the problems we want to explore next. Training data is critical for the accuracy of convolutional neural network, and a natural idea is to collect more suitable data from various situations especially in the driving scenes.

REFERENCES

[1] M. Q. Khan and S. Lee, "A comprehensive survey of driving monitoring and assistance systems," *Sensors*, vol. 19, no. 11, 2019.

[2] H. A. Khojasteh, A. A. Alipour, E. Ansari, and P. Razzaghi, "An intelligent safety system for human-centered semi-autonomous vehicles," 2018.

[3] L. R. Hartley, "Fatigue and driving: Driver impairment, driver fatigue, and driving simulation," 1995.

[4] M. Zahra, A. S. N. Miri, and M. Mohammad, "Eeg-based drowsiness detection for safe driving using chaotic features and statistical tests," *Journal of Medical Signals & Sensors*, vol. 1, no. 2, pp. 130–137, 2011.

[5] M. N. Husen, S. Lee, and M. Q. Khan, "Syntactic pattern recognition of car driving behavior detection," in *International Conference on Ubiquitous Information Management & Communication*, 2017.

[6] H. A. Khojasteh, A. A. Alipour, E. Ansari, and P. Razzaghi, "An intelligent safety system for human-centered semi-autonomous vehicles," 2018.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015.

[9] J. Dai, L. Yi, K. He, and S. Jian, "R-fcn: Object detection via region-based fully convolutional networks," 2016.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2018.

[11] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016.

[12] Intel, "Openvino toolkit." https://software.intel.com/en-us/openvino-toolkit.

[13] Intel, "Open model zoo repository." https://github.com/opencv/open_model_zoo.

[14] J. Huang, V. Rathod, S. Chen, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, S. Yang, and S. Guadarrama, "Speed/accuracy trade-offs for modern convolutional object detectors," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," 2017.

[16] X. Guo, S. Li, J. Zhang, J. Ma, and H. Ling, "Pfld: A practical facial landmark detector," 2019.

[17] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] A. Jourabloo, Y. Mao, X. Liu, and R. Liu, "Pose-invariant face alignment with a single cnn," 2017.

[19] S. Xiao, J. Feng, L. Liu, X. Nie, and A. Kassim, "Recurrent 3d-2d dual learning for large-pose facial landmark detection," in *IEEE International Conference on Computer Vision*, 2017.