

Night Fatigue Driving Detection Technology Using Infrared Images and Convolutional Neural Networks

Huei-Yung Lin¹ ^a and Kai-Chun Tu²

¹*Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 106, Taiwan*

²*Department of Electrical Engineering, National Chung Cheng University, Chiayi 621, Taiwan*

Keywords: Fatigue Driving Detection, IR Images, Convolutional Neural Network, Action Recognition.

Abstract: Traffic accident is one of top ten causes of death, and fatigue driving is one of the major reasons. It usually reduces the driver's concentration and reaction speed, and is especially dangerous in some situations at night. This work presents a real-time driving fatigue monitoring system. The proposed network architecture with Unbalanced Local CNNs can effectively draw attentions to different face regions according to driver's states due to fatigue. Based on SlowFast, the recognition accuracy of our method on the IR image datasets is greatly improved compared to the original model. Moreover, an adversarial learning mechanism is incorporated to extract the common features of daytime RGB and nighttime IR images to increase the overall robustness. The experiments carried out on public datasets and road scene images have demonstrated the effectiveness of the proposed technique. The code is available at <https://github.com/KaiChun-Tu/slowfastDrowsyDriver>


1 INTRODUCTION

According to the NHTSA (National Highway Traffic Safety Administration), fatigue driving is a major cause of traffic accidents in the United States. There were 697 people died due to fatigue driving in 2019, and 2.5% of fatal traffic accidents and 2.0% of non-fatal traffic accidents were related to drowsy driving. Between 2009 and 2010, the CDC (Centers for Disease Control and Prevention) interviewed a total of 147,076 people in 19 states and the District of Columbia, 4.2% of respondents admitted that they have once fallen asleep while driving. NHTSA's survey on the number of car accidents in age groups at different times of the day revealed that drivers under the age of 45 have far more car accidents between 11 p.m. and 8 a.m. than other car accidents happened in other time.

The above information clearly shows that fatigue driving is a very dangerous situation. Although the driver might be able to hold back and does not fall asleep, fatigue will still reduce the driver's reaction speed, concentration, decision-making ability, and increase the chance of car accidents. Thus, fatigue driving is a serious traffic problem and needs to be solved urgently (Zhang and Lin, 2021).

Since the introduction of AlexNet (Krizhevsky et al., 2012), deep neural networks have been developed rapidly and used in various transportation fields. Deep learning approaches are applied to driving fatigue detection and technical advances have been reported (Sikander and Anwar, 2018). However, most algorithms take RGB images captured during the day as input for network training, and cannot be generalized to nighttime application scenarios. Since fatigue driving is more likely to occur at night, it is necessary to develop a system exclusively for nighttime use. In addition, current methods consider the task as an image recognition problem, and annotate training data frame-by-frame for fatigue driving detection. It might have some difficulties to distinguish the behaviors of falling asleep and blinking when only a single image is used.

In this work, we develop a nighttime driver fatigue detection system using the images acquired by an infrared camera. The proposed network structures take IR images as input, and model the problem as abnormal behavior classification from an image sequence. We collect both daytime RGB images and nighttime IR images for training and testing. A network architecture based on multiple LANets is combined with SE-blocks to focus on different areas according to different fatigue actions. Moreover, a modality-invariant

^a  <https://orcid.org/0000-0002-6476-6625>

feature extraction module based on adversarial learning is added to improve the robustness of fatigue detection by an image type dependent discriminator. In the experiments, ablation studies on real scene images are carried out to verify the feasibility and effectiveness of the proposed technique.

2 RELATED WORK

In the existing literature, the detection of drowsy driving is based on two types of information sources, including biomedical signal sensing and external characteristics extraction (Doudou et al., 2020). Biomedical signals refer to the electric currents generated by the human body. In recent works, there are three types of signals commonly used to distinguish the states of fatigue, namely EEG (electroencephalogram), ECG (electrocardiogram) and EOG (electrooculogram).

EEG is the signal produced by measuring the potential differences between groups of cerebral cortex cells. In an early investigation, (Jap et al., 2009) collected the frequencies of four EEGs in a fatigued driving state, and found that the signals from some areas of a brain changed obviously. Since the brain activity is directly related to fatigue, EEG signals can effectively reflect the state of fatigue and provide accurate results. However, EEG is not only related to fatigue, but also arms, eyeballs, mouth, etc. Many parts of human body are controlled by the brain, so EEG is very susceptible to other noise. Thus, it is required to carry out an additional preprocessing stage.

ECG utilizes electrodes to detect and amplify the signals produced by a very small potential change in the skin caused by the heartbeat. To measure the ECG signals during driving, early approaches placed electrodes in the seat belt (Murugan et al., 2020). Nevertheless, this can be easily influenced by the dynamic driving environment. In a later implementation (Warnecke et al., 2022), the electrodes are attached to the steering wheel. But this approach still has the shortcomings of low accuracy, and is easily to be affected by the noise.

EOG records the voltages between the retina and the cornea of the eye. This kind of signals needs to be measured by electrodes attached to both sides of the eyes, and therefore have a serious impact on normal driving. To cope with this problem, some techniques proposed to place the electrodes on the forehead of the driver (Zhang et al., 2015). However, the applicability is still very limited in practical uses.

For fatigue or drowsy driving detection, extrinsic characteristics refer to the visible appearance features related to the driver's state, such as nodding, yawning,

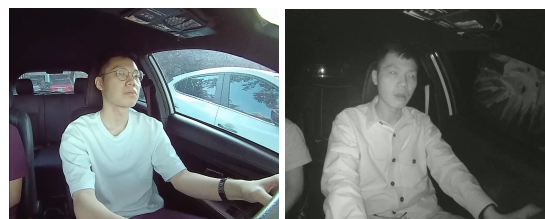


Figure 1: The dataset images captured in the daytime (left) and nighttime (right).

blinking frequency (Dong and Lin, 2021). Considering the fatigue characteristics related to the eyes, it is commonly evaluated by the duration of eye closed, or its proportion per unit time. In addition, the mouth also represents an important factor for the fatigue condition. There are many techniques combining the features extracted from eyes and mouth (Savaş and Bericikli, 2020). These approaches first utilize convolutional neural networks to identify the eyes and mouth. It is then followed by the eyes closing time per unit period and the aspect ratio of the mouth to evaluate the fatigue state.

In addition to the use of a single image for fatigue driving detection, utilizing image sequences with the temporal information can provide more robust results. It is usually difficult to distinguish the situations such as blinking and falling asleep or yawning and talking with a still image. Thus, several recent techniques are developed based on optical flow, LSTM and 3DCNN (Quddus et al., 2021). The optical flow obtained using two consecutive images can provide the movement of features to improve the recognition rate. For 3DCNN, an additional sliding window is added to 2DCNN to acquire the temporal information between adjacent images. In LSTM, long short-term memory is adopted as a specialized recurrent neural network. It considers the features from neural networks integrated along the time axis.

There are only a limited number of datasets available for the fatigue driving evaluation. In the YawDD dataset, the images were captured under the natural daylight (Abtahi et al., 2020). It consists of four action states for the drivers: talking, singing, silence and yawning. The dataset was collected with 47 males and 43 females from different regions and under different ages. It provides only the RGB images and cannot be used to train the neural networks for nighttime fatigue detection. Furthermore, there are only mouth features provided by YawDD. Without the information related to eyes and head, the detection of certain behaviors, such as falling asleep and nodding, will become more challenging. Another commonly used public dataset is the DDDD dataset (Weng et al., 2016). The images were captured using an IR camera in an indoor envi-

ronment under daytime and nighttime lighting conditions. Each driver was filmed in five scenarios: wearing sunglasses in daytime, not wearing glasses in daytime, wearing glasses in daytime, not wearing glasses in nighttime, and wearing glasses in nighttime. Since the DDDD images were collected indoors, it does not reflect the actual lighting and environment illumination changes in the outdoor scenes.

3 METHOD

It is indicated by NHTSA that most car accidents happened in the nighttime, but a majority of existing fatigue detection research was conducted based on RGB images. Due to the lack of ambient light in the night scenes, it is not feasible to adopt the RGB images acquired at night for fatigue detection. In this work, we use IR images captured at night for processing. These images are less likely to have the reflection issue of glasses compared to the daytime RGB images. They are also better suited for the networks to have attentions on the eyes and mouth. Thus, we propose a convolutional neural network which emphasizes the use of IR characteristics of images.

To collect nighttime IR images, we used a Garmin Dash Cam Tandern camera. It is equipped with an active infrared function to capture clear driving images in the absence of light sources during the night. Our dataset was acquired in order to match the actual application scenarios as much as possible. There were 8 different drivers in a vehicle with no light source other than streetlights filmed for about four minutes in both fatigue and normal states. For the testing purpose, we also collected image data during the day. The camera automatically switch between the IR and RGB modes according to the ambient light. Figure 1 shows examples of RGB and IR images acquired in the daytime and nighttime, respectively.

To label the ground-truth data, we adopt a method different from the frame-by-frame approach for most public datasets. In our technique, the fatigue detection task is considered as finding abnormal events, and the image annotation is carried out the same way as for the Kinetics dataset (Carreira and Zisserman, 2017). The image sequence is marked in groups of 8 frames, with 3 images shared between the consecutive groups, to generate distinct samples while maintaining overall data volume.

In the proposed method, fatigue detection is based on the facial features of the driver, such as eye closing, yawning, head tilting or nodding, etc. Thus, face images are first extracted using RetinaFace (Deng et al., 2020) to generate the input sequences for fatigue driv-

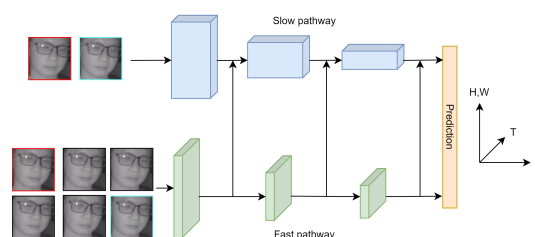


Figure 2: The network architecture of SlowFast modified for this work.

ing detection. The backbone of our proposed network architecture is based on the SlowFast model (Feichtenhofer et al., 2019). Different from the 3D convolution models utilizing the same spatial and time dimensions, SlowFast considers the temporal information changing very rapidly compared to the spatial semantic information. That is, it is not possible that one frame with yawning, the next frame with talking, and then followed by the third frame with yawning again. Since the facial expression changes rapidly, SlowFast establishes two separate paths to extract the spatial semantics and temporal action information.

3.1 SlowFast with Attention Module

Figure 2 depicts the SlowFast architecture adopted in this work. The slow pathway has image input at a lower frame rate. It contains a larger amount of parameters in the channel dimension for the extraction of spatial semantic information. In contrast, the fast pathway inputs images at a higher frame rate in the time dimension. The larger amount of parameters in the time dimension is used to extract the motion information. By combining these two pathways, SlowFast connects the motion information to the Slow pathway for data fusion at each stage.

The main difference between IR and RGB images is that the ambient lights at night are much less compared to the daytime. Thus, the reflective brightness on the glasses and appeared in the image causes a serious impact on fatigue detection. Most existing techniques extract the eyes and mouth to improve the accuracy. However, this method can be greatly affected if the glasses reflect lights and cause the loss of eye features. Since IR images record much less reflective appearance on the glasses, they are more suitable for the approach. In addition to the extra efforts for the extraction of eyes and mouth regions, the network structure also increases in the model size and processing time

The above issues are generally very unfavorable for the practical use. To deal with these problems, the proposed method utilizes face images without further extraction of eye and mouth regions. Alternatively, an

additional module is added to the architecture so that the network model is able to focus more on eyes and mouth. In our network structure, the SlowFast model is adopted. Since it contains the information along the time axis, which is not necessary for current attention models taken as input, we utilize the average pooling operation to compress the temporal information of the feature maps instead.

3.2 Unbalanced Local CNNs

In recent years, there has been a mainstream research on the facial expression recognition. It is very similar to the objective of this work in the input/output (i.e., image/category) modeling of the system. Hence, it is desirable to adopt the method to improve the network performance. More specifically, the adjustment of facial regions for expression recognition can be used to provide the attention for the network. In the previous work, an attention adjustment model, Local CNNs, is proposed (Xue et al., 2021). The basic idea is to first use multiple LANets (Ding et al., 2020) to generate attention maps which focus on different regions. It is then followed by combining all attention maps via finding the maximum value of each pixel in the attention map. The MAD module in the architecture is utilized to prompt the network to explore more potentially important areas similar to dropout by randomly reducing the attention maps to zero.

For facial expression recognition, the image areas involved in each expression are different, and attention maps generated by Local CNNs need to be averaged and cannot be too imbalanced. However, in our driving fatigue detection, the network only has to pay attention to the eyes and mouth. It is not necessary to care about the imbalance of generated attention maps, since the important parts for fatigue may be different each time. For example, we expect the network model being able to focus exclusively on the mouth features when yawning.

With the modification to the original model, the MAD module is removed and an additional fully connected layer is added to utilize the input features to calculate the importance of the attention maps generated by each LANet. The original Local CNNs tend to result in a more average attention map since they simply combine the maximum value of each attention map. Our Local CNNs can make one or two attention maps completely larger than the others based on the features computed by fully connected layers. Consequently, the final attention map has only the features of fewer attention maps, and provide a less averaged result. Thus, this architecture is referred to as Unbalanced Local CNNs.

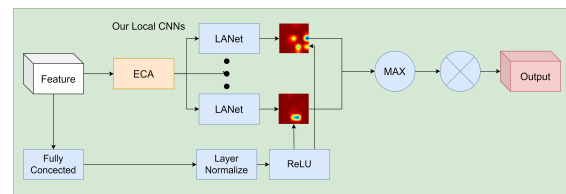


Figure 3: Our final Local CNNs model.

In our original design, we inserted the Unbalanced Local CNNs in the slow pathway which is responsible for spatial operations. It is expected to have the spatial semantic information focused on the positions such as the eyes and the mouth. Due to the convolution operation of previous layers, the features have been focused on uninteresting areas. When the fifth layer is added to our Local CNNs, the attention of the model cannot be effectively adjusted. Since the eyes and mouth are expected to have larger movements than other regions and the fast pathway is responsible for motion detection, the features of the fast pathway are used instead as the input for attention map generation. It is then used to adjust the attention of the spatial semantics of the slow pathway to improve the overall performance.

In a recent study (Woo et al., 2018), CBAM (Convolutional Block Attention Module) was proposed to make a joint attention adjustment through the channel and spatial attention modules. The network structure of our Local CNNs is also a spatial attention module, and a channel attention module is incorporated ahead. By adding the new ECA (Efficient Channel Attention) module (Wang et al., 2020), the final architecture of our Unbalanced Local CNNs is shown in Figure 3.

Figure 4 shows the attention maps of each model. The first column is the input image, and the second to fourth columns are attention maps of SlowFast+our Local CNNs, SlowFast, and SlowFast+Local CNNs, respectively. It can be observed that the attention map of SlowFast+Local CNNs is quite messy and without specific focus. The attention of SlowFast is more concentrated, but it also focuses on other parts of the face besides the eyes and mouth. Our network model will focus on different aspects of the driving status. It constantly observes the eye features, but the mouth would receive the highest attention when yawning. Although yawning will also cause changes in the entire face, the attention is comparably less significant.

3.3 Modality Invariant Feature Learning

In general, nighttime and daytime images have very different characteristics. Even both cases are recorded with IR images, it is still difficult for a single network

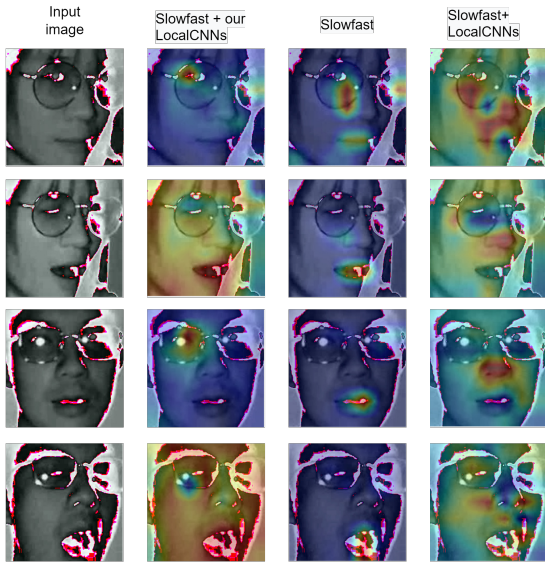


Figure 4: The heatmaps of each network model.

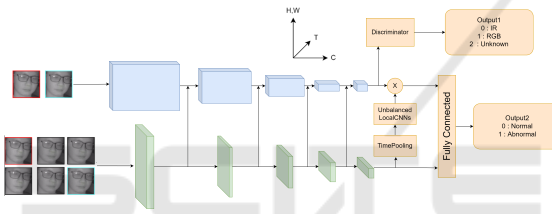


Figure 5: The modality-invariant feature learning based on adversarial learning.

to derive good results for day and night scenes simultaneously. The performance will be further degraded if IR images are used for nighttime and RGB images are used for daytime. Therefore, most of the current fatigue detection techniques only work on either daytime or nighttime images. In the recent research on pedestrian recognition, the accuracy of the network model is improved by learning the modality-invariant features for pedestrian re-identification between RGB and IR images. Based on a similar idea, we adopt the adversarial learning utilizing modality-invariant features (Lin et al., 2022) to construct our fatigue detection model, and the network architecture is shown in Figure 5.

In the network modification, we insert a discriminator to the fifth layer of SlowFast. This discriminator is responsible for determining whether the input features are from IR images, RGB images, or modality-invariant. Two loss functions

$$Loss_D = CE(out_D, Label_I) \quad (1)$$

and

$$Loss_E = CE(out_D, Label_M) \quad (2)$$

Table 1: Ablation study of the proposed network model. A sequence of 8 IR images is used for evaluation.

| | F1 score |
|----------------------------------|----------|
| SlowFast | 0.954 |
| SlowFast + Attention Augmented | 0.960 |
| SlowFast + Local CNNs | 0.961 |
| SlowFast + Unbalanced Local CNNs | 0.965 |

are used in the proposed network, where CE is cross-entropy and out_D represents the output of the discriminator. In Eq. (1), $Label_I$ indicates the true category of the image, and $Loss_D$ represents the loss between the discriminator output and real class. The discriminator will be optimized according to the back-propagation gradient of $Loss_D$ to enhance its ability to distinguish the input features from IR or RGB images. In Eq. (2), $Label_M$ represents the category with modality invariant, and $Loss_E$ denotes the loss between the discriminator output and modality-invariant class. The first to the fifth convolutional layers in SlowFast is optimized according to the back-propagation gradient of $Loss_E$ to improve the ability of feature extraction from RGB and IR images.

By the adversarial learning established using the discriminator and two loss functions in SlowFast, the feature extraction capability can be improved by confrontation. We can effectively make the convolutional layers work on both IR and RGB image inputs. Common features will be extracted, so the fatigue state can be accurately identified in the subsequent fully connected layers.

4 EXPERIMENTS

To ensure the effectiveness of our proposed model, we first perform several ablation experiments, including spatial attention at night, temporal feature compression, and channel attention. We use IR images to test the original SlowFast model with the following attention models, Attention Augmented, Local CNNs, and Unbalanced Local CNNs (without channel attention module), and the performance is tabulated in Table 1. The results indicate that SlowFast + LocalCNNs improves the F1 score by 0.7% for night scenes. Our Unbalanced Local CNNs is 1.1% higher than the original SlowFast and 0.4% higher than SlowFast + Local CNNs. Moreover, the parameters of the proposed model are only about 40% of SlowFast + Attention Augmented. The experiments have demonstrated that our method has good performance in terms of accuracy, processing speed, and model parameters for IR image input.

For channel attention, we test the following four models, SE-block, CBAM, Coordinate Attention, and

Table 2: Ablation study of different channel attention modules.

| | F1 score | FPS | Model Size |
|----------------------|----------|-------|------------|
| SE-block | 0.966 | 78.62 | 136.8 MB |
| CBAM | 0.953 | 76.75 | 136.8 MB |
| Coordinate Attention | 0.935 | 77.19 | 136.8 MB |
| ECA | 0.969 | 79.38 | 136.8 MB |

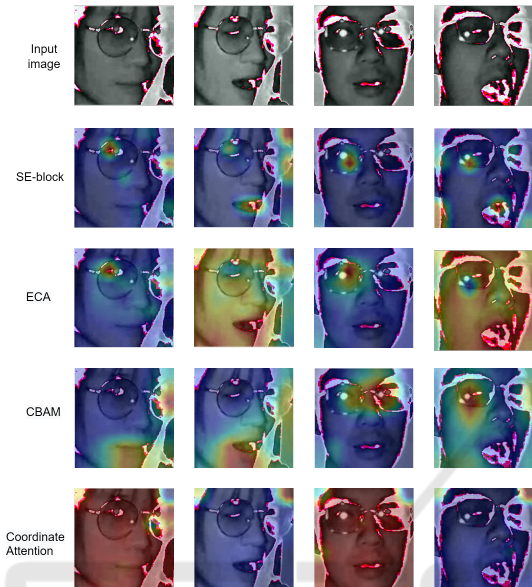


Figure 6: Attention maps of different channel attention modules.

ECA. Table 2 tabulates the evaluation results. It can be seen that the performance of using ECA is slightly higher using SE-block, but there is a large difference between CBAM and Coordinate Attention, with 3.4% higher for CBAM. The attention maps derived from different channel attention models are shown in Figure 6. The results from the SE-block provide almost no additional ROIs other than the eyes and mouth, and only focus on the mouth when driving and yawning. Nevertheless, the attention of the mouth is difficult to accurately cover the entire mouth region. For CBAM, some face images are focused on the mouth area and others are focused on the eyes. As for Coordinate Attention, the overall attention distribution is quite average, and there is no place to pay more attention.

For feature compression in the time dimension, we evaluate three approaches in compressing spatial features from CBAM (Woo et al., 2018), average pooling, max pooling, and average pooling+max pooling, and the F1 scores are 0.969, 0.958 and 0.963, respectively. Finally, we compare our method with the work (Liu et al., 2019). Taking 8 IR images as input, the F1 scores of the proposed SlowFast + Unbalanced Local CNNs and the baseline approach (Liu et al., 2019) are 0.969 and 0.941, respectively.

Table 3: The experimental results of the IR images in NTHU-DDDD dataset.

| | F1 score |
|-----------------------------------|----------|
| SlowFast + our Local CNNs (train) | 0.96 |
| SlowFast + our Local CNNs | 0.73 |
| (Bai et al., 2021) | 0.854 |
| (Lyu et al., 2018) | 0.9005 |
| (Park et al., 2016) | 0.748 |
| (Yu et al., 2016) | 0.683 |
| (Liu et al., 2019) | 0.962 |

Table 4: The experimental results using YawDD dataset.

| | F1 score |
|---------------------------|----------|
| SlowFast + our Local CNNs | 0.970 |
| (Bai et al., 2021) | 0.895 |
| (You et al., 2020) | 0.943 |

4.1 Public Datasets

The proposed fatigue driving detection technique is tested on two public datasets, NTHU-DDDD (Weng et al., 2016) and YawDD (Abtahi et al., 2014). In the DDDD dataset, the images are marked in a way that one image corresponds to one category. The annotation is modified to conform our network input format in a group of eight images. Table 3 shows the results evaluated using IR images and compared with several different algorithms (Bai et al., 2021; Lyu et al., 2018; Park et al., 2016; Liu et al., 2019). It can be seen that the F1 score of our model is 73%, but the performance during training is 96%. The serious overfitting problem might be caused by the data labeling with a long period of time instead of image by image. As a result, there will be abnormal images very similar to normal images, no matter for training or testing images. Our model performs poorly due to the learnable features in training data not extended to the testing images.

In the YawDD dataset, there are three types of annotations: Normal, Talking and Yawning. It is labeled for video clips of about 20 to 40 seconds. Since it is not possible to have driving and yawning for the entire video clips, there are Normal images mixed in the Yawning category images. Thus, we use the annotation file for fine-grained classification of videos in the Yawning category. It is also adopted for data segmentation of Yawning category videos. For Normal and Talking videos, we randomly select 14 drivers and divide them into training and testing data. The evaluation results and comparison with other techniques are shown in Table 4.

4.2 Dynamic Images and RGB/IR Mix

In the previous works, most experiments were carried out in indoor environments or using images captured

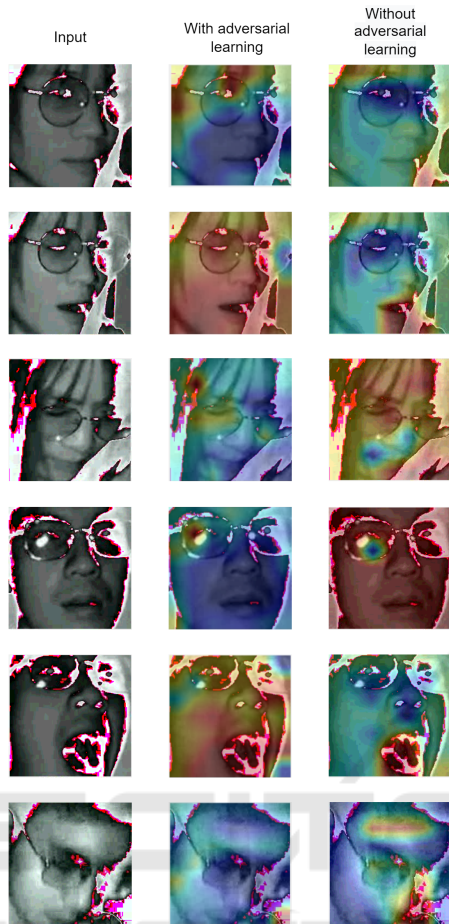


Figure 7: The attention maps derived using the mixture of IR and RGB images.

in a stationary vehicle. However, in practical applications, the fatigue detection system will operate under the vehicle in motion. The current approaches usually extract the face regions, so the impact of environment illumination changes should be minimized. To further investigate the influences of real driving to the fatigue detection, we perform the experiment using a driving video captured at night. The test result shows the F1 score of 94.1%, about 2.8% lower than the static case. Same as the previous observation, the false detection mostly happened during the transition between normal and abnormal states.

In addition to network training and testing both on nighttime IR images, it is also desirable to make our technique work on daytime application scenario using RGB images. Thus, in the last experiment we mix IR images recorded at night with RGB images captured during the day using the same hardware settings for training. The network model is tested for both daytime and nighttime scenes, and the results are shown in Table 5. It can be seen that contrastive learning can

Table 5: The evaluation results using the mixture of IR and RGB images.

| | F1 score | FPS | Model Size |
|---------------|----------|-------|------------|
| Without A. L. | 0.941 | 77.60 | 135.1 MB |
| With A. L. | 0.956 | 72.29 | 136.8 MB |

effectively improve the recognition result, and the accuracy is about 1.5% higher. Since the discriminator does not need to perform operations during inference, the overall FPS does not change too much. Figure 7 shows the attention maps derived from this model. By adding adversarial learning, the attention can be effectively focused on the eyes without special actions. In the yawning region, the mouth shows good attention with or without adversarial learning. When the driver is nodding, however, neither cases provide a particular area of interest.

5 CONCLUSIONS

In this paper, we develop a nighttime fatigue driving detection technique using infrared images. Since the commonly used public datasets for fatigue driving research provide indoor or RGB images, there are many limitations for nighttime scenes. This work uses an IR camera to collect fatigue driving images in the actual application scenario. We propose a network architecture which is able to effectively use action information in the SlowFast model to change the spatial attention. With limited parameters and computing time, the network can focus on the eyes and mouth regions. Moreover, a modality invariant feature learning mechanism based on adversarial learning is added to improve the accuracy for both the daytime and nighttime scenes. The experiments carried out on real application scenarios have demonstrated the effectiveness of the proposed technique.

ACKNOWLEDGMENTS

This work was financially/partially supported by the Ministry of Science and Technology of Taiwan under Grant MOST 109-2221-E-027-126-MY3 and Create Electronic Optical Co., LTD, Taiwan.

REFERENCES

- Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., and Hariri, B. (2014). Yawdd: A yawning detection dataset. In *Proceedings of the 5th ACM multimedia systems conference*, pages 24–28.

- Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., and Hariri, B. (2020). Yawdd: Yawning detection dataset.
- Bai, J., Yu, W., Xiao, Z., Havyarimana, V., Regan, A. C., Jiang, H., and Jiao, L. (2021). Two-stream spatial-temporal graph convolutional networks for driver drowsiness detection. *IEEE Transactions on Cybernetics*, pages 1–13.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.
- Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ding, L., Tang, H., and Bruzzone, L. (2020). Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):426–435.
- Dong, B.-T. and Lin, H.-Y. (2021). An on-board monitoring system for driving fatigue and distraction detection. In *2021 22nd IEEE International Conference on Industrial Technology (ICIT)*, volume 1, pages 850–855. IEEE.
- Doudou, M., Bouabdallah, A., and Berge-Cherfaoui, V. (2020). Driver drowsiness measurement technologies: Current research, market solutions, and challenges. *International Journal of Intelligent Transportation Systems Research*, 18(2):297–319.
- Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.
- Jap, B. T., Lal, S., Fischer, P., and Bekiaris, E. (2009). Using eeg spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2):2352–2359.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- Lin, X., Li, J., Ma, Z., Li, H., Li, S., Xu, K., Lu, G., and Zhang, D. (2022). Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20973–20982.
- Liu, W., Qian, J., Yao, Z., Jiao, X., and Pan, J. (2019). Convolutional two-stream network using multi-facial feature fusion for driver fatigue detection. *Future Internet*, 11(5):115.
- Lyu, J., Yuan, Z., and Chen, D. (2018). Long-term multi-granularity deep framework for driver drowsiness detection. *arXiv preprint arXiv:1801.02325*.
- Murugan, S., Selvaraj, J., and Sahayadhas, A. (2020). Detection and analysis: driver state with electrocardiogram (ecg). *Physical and engineering sciences in medicine*, 43(2):525–537.
- Park, S., Pan, F., Kang, S., and Yoo, C. D. (2016). Driver drowsiness detection system based on feature representation learning using various deep networks. In *Asian Conference on Computer Vision*, pages 154–164. Springer.
- Quddus, A., Zandi, A. S., Prest, L., and Comeau, F. J. (2021). Using long short term memory and convolutional neural networks for driver drowsiness detection. *Accident Analysis & Prevention*, 156:106107.
- Savaş, B. K. and Becerikli, Y. (2020). Real time driver fatigue detection system based on multi-task connn. *Ieee Access*, 8:12491–12498.
- Sikander, G. and Anwar, S. (2018). Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2339–2352.
- Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539.
- Warnecke, J. M., Ganapathy, N., Koch, E., Dietzel, A., Flormann, M., Henze, R., and Deserno, T. M. (2022). Printed and flexible eeg electrodes attached to the steering wheel for continuous health monitoring during driving. *Sensors*, 22(11):4198.
- Weng, C.-H., Lai, Y.-H., and Lai, S.-H. (2016). Driver drowsiness detection via a hierarchical temporal deep belief network. In *Asian Conference on Computer Vision*, pages 117–133. Springer.
- Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.
- Xue, F., Wang, Q., and Guo, G. (2021). Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610.
- You, F., Gong, Y., Tu, H., Liang, J., and Wang, H. (2020). A fatigue driving detection algorithm based on facial motion information entropy. *Journal of advanced transportation*, 2020.
- Yu, J., Park, S., Lee, S., and Jeon, M. (2016). Representation learning, scene understanding, and feature fusion for drowsiness detection. In *Asian Conference on Computer Vision*, pages 165–177. Springer.
- Zhang, J.-Z. and Lin, H.-Y. (2021). Driving behavior analysis and traffic improvement using onboard sensor data and geographic information. In *VEHITS*, pages 284–291.
- Zhang, Y.-F., Gao, X.-Y., Zhu, J.-Y., Zheng, W.-L., and Lu, B.-L. (2015). A novel approach to driving fatigue detection using forehead eog. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 707–710. IEEE.