A Driver Distraction Detection Method Based on Convolutional Neural Network

2022 7th International Conference on Computer and Communication Systems (ICCCS) | 978-1-6654-5060-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCCS55155.2022.9846062

Chuheng Wei*
Warwick Manufacturing Group
The University of Warwick
Coventry, UK
chuheng.wei@warwick.ac.uk

Chuanshi Liu
School of Mechanical and Aerospace Engineering
Jilin University
Changchun, China
liucs9920@mails.jlu.edu.cn

Shaocui Chi
School of Information and Communication Engineering
Communication University of China
Beijing, China
chishaocui@cuc.edu

***Abstract*—The growth of the economy and technology is increasing the popularity of automotive, but it also increases the number of traffic accidents. The driver's factor is a major cause of traffic accidents and ensuring the driver's concentration while driving is an essential research topic along with the development of autonomous cars. Recent developments in artificial intelligence and advanced hardware systems have made convolutional neural networks increasingly useful in computer vision. The purpose of this article is to explore the use of ResNet-50 neural networks in detecting driver distractions. In this article, the performance of ResNet-50 neural network is studied and analyzed and the possibility of its use for distraction detection is explored. In addition, it is found that this neural network is more capable of classifying whether a driver is distracted than of classifying their specific distracted behavior.**

Keywords—driver distracted detection, ResNet-50, Convolution neural network, classification

I. INTRODUCTION

With the advancement of science and technology, people's living standards improved, which led to an increase in the number of cars over the last few decades [1]. As a result, an increasing number of people obtain drivers' licenses. On the other hand, with the rapid expansion of traffic scale, the number of traffic accidents has also increased significantly. There is an increase in the availability of more sophisticated automotive infotainment systems that are designed to enhance the driver's driving experience. However, these systems also cause visual, biomechanical, and cognitive distractions and may affect driving performance in different ways [2]. According to the data of the NHTSA (National Highway Traffic Safety Administration) [3], driver distraction is an important cause of traffic accidents, especially rear end crashes. In recent years, the development of computer chip technology and the development of artificial intelligence methods have led to the development of driver health monitoring systems in many vehicles [4]. Numerous renowned automotive manufacturers, government agencies, and research institutions have conducted research on related topics [5].

A driver monitoring system based on ResNet deep neural networks is presented in this paper, and its performance is evaluated and analyzed. The following chapters of this paper address:

Section II introduces prior knowledge of driver distraction detection and deep neural networks. Section III describes the methodology used for the experiments. Section IV analyzes the experiments and presents the results of the experiments. Section V summarizes the experimental results and discusses possible future developments.

II. PRIOR KNOWLEDGE

A. Distraction

What is considered a driver state is not universally defined and is often used loosely by psychologists and engineers alike. Gonçalves and Bengler [6] describes the driver state as a set of conditions that affect an individual driver. An active and focused driver is able to deal with the conditions encountered on the road, whereas a driver who is distracted is significantly impaired in his or her ability to control the vehicle [7] . Data indicate that more than half of the crashes involving inattention are caused by driver distraction [3], and the causes of driver distraction vary widely. Lee et al. [8] presents a generally accepted definition of driver distraction based on the extensive study of various definitions in the literature: “Driver distraction is a diversion of attention away from activities critical for safe driving toward a competing activity.”

B. Driver Distraction Detection

Distracted driving behavior is one of the leading causes of traffic accidents. Numerous experts, scholars, and scientific research institutions have contributed to the field of detecting driver distractions[9, 10]. Driver distraction detection generally starts from five indicators: eye movement information, physiological signals, subjective evaluation, vehicle environment information, and a mixed indicator of the fusion of several indicators [11-13].

There are two types of detection methods available for detecting driver distraction: contact sensor detection and non-contact sensor detection. Firstly, as the name implies, the contact-sensing, detection method attaches or wears the sensor device to the driver's body, and judges the driver's driving posture by detecting the driver's physiological signal indicators such as brain electricity, electrocardiogram, skin electricity and muscle electricity [14]. Data from the devices can be used to determine whether the driver is distracted, which has a high

degree of accuracy[15]. Secondly, another method is the non-contact sensor detection technique, which uses pressure sensors, proximity sensors, and cameras to monitor a driver's movements and inspect vehicle environment, subjective evaluation and other indicators of driver distraction [2]. The algorithm presented in this paper selects the second type of method, which uses deep neural networks to analyze visual information to determine the state of the driver.

C. *Convolutional Neural Network*

Convolutional neural networks are feed-forward neural networks with deep neural network structures that utilize convolutional operations[16]. On the basis of biological visual perception systems, convolutional neural networks are constructed. In addition, the convolutional neural network is a type of artificial neural network whose design is based on ways animals communicate[17]. It is composed of large numbers of artificial neurons that work together to form a network system.

In general, a convolution neural network contains a layer of convolution and a layer of pooling. In the case of processing the input image, the convolution neural network may replace the cumbersome process of feature extraction and data reconstruction in the previous algorithm[18]. Through multiple convolution layers and pooling layers, the input image is processed through feature extraction, and gradually changes from low-level to high-level features, allowing the image to be directly input as data without further processing.

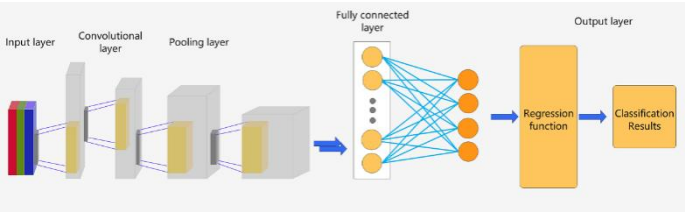


Fig. 1. The structure of Convolutional Neural Network (CNN).

Due to the unique advantages of convolutional neural networks for image processing, convolutional neural networks have been widely used in computer vision fields such as image recognition, target detection, and pose estimation[16]. Figure 1 depicts a convolutional neural network model. Convolutional neural networks generally consist of five layers, namely input layer, convolutional layer, pooling layer, fully connected layer, and output layer.

III. METHODOLOGY

ResNet is the abbreviation of Residual Network, which is a classic model for convolutional neural networks. It was proposed by Dr. He Yuming et al. [19] of Microsoft Research, which successfully trained a 152-layer neural network by using residual units and won the championship in the ILSVRC2015 competition [20]. ResNet has a deep level of abstraction, which reduces the amount of computational parameters and has a highly pronounced effect [21]. In the field of object classification and classical neural networks as the backbone of computer vision tasks, the ResNet family of networks has become an immensely important component[22].

Based on ResNet-50, this paper builds a model for detecting driving distraction. The model identifies driving distraction by detecting the driver's actions and facial features from the frame of the RGB video. The deep ResNet50 convolutional neural network is used to ensure the accuracy of the driving image fatigue detection model in this paper. The gradient explosion or gradient disappearance may occur during the training process as the number of layers of the network increases, causing the training network to fail to converge. Gradient explosion is caused by the fact that convolutional neural networks update their parameters by backpropagation, and that the chain rule of backpropagation multiplies multiple gradients to derive the gradient of the underlying module, and when these gradients contain more than one value, it will result in gradient explosion. Similarly, when these gradients contain multiple values less than 1, the gradients will disappear. ResNet is designed to allow the gradient to circulate efficiently by designing the residual structure. The residual structure is shown in Figure 2.

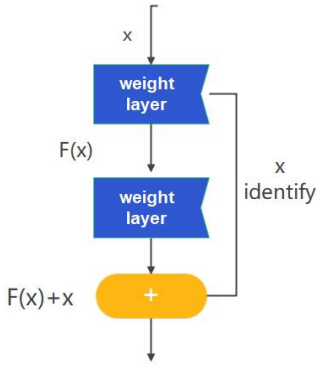


Fig. 2. The residual structure of ResNet.

This residual structure can be specifically interpreted as:

The learned feature in the stacked layer structure is denoted as $H(x)$ when the input is x, and the actual original learned feature is $F(x)+x$ when it is expected to learn the residual $F(x)=H(x)-x$.

Residual learning is easier than learning directly from the original features. When the residual is 0, the stacking layer is only doing constant mapping at this point, so network performance does not degrade. However, the residuals are not actually zero, which also allows the stacking layer to learn new features on top of the input features and thus have better performance. The structure of residual learning is shown in Figure 2. This short-circuits-link like structure allows the shallow module to directly participate in the deep module by jumping the connection. Therefore, the gradient can be directly transferred from the deep module to the shallow module during the backpropagation process, which effectively alleviates the gradient flow problem.

The workload is greatly reduced through residual learning, and the specific analysis process is as follows

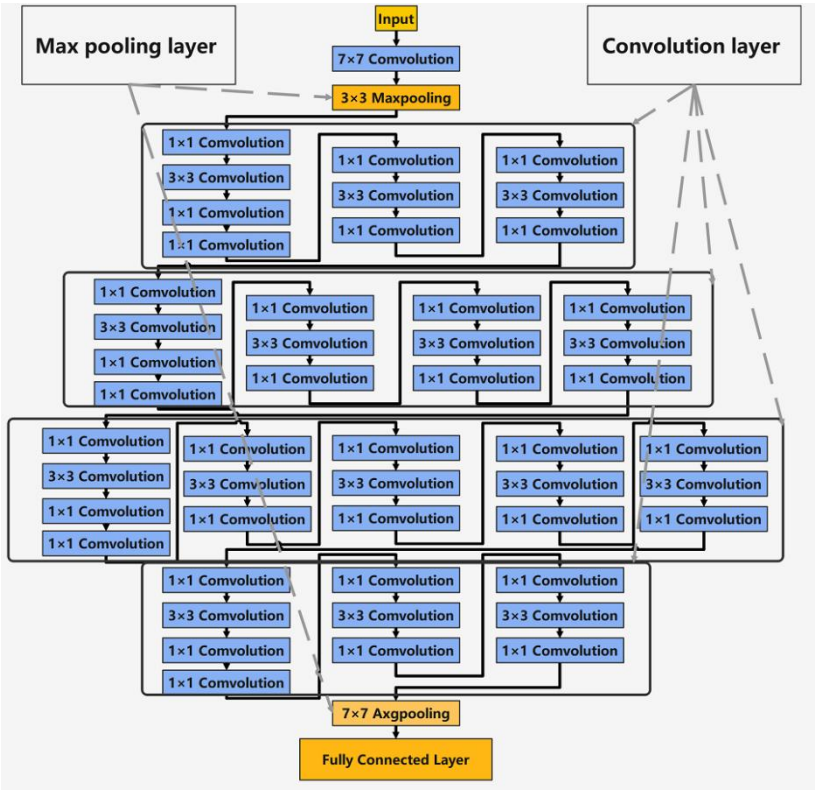


Fig. 3. The architecture of ResNet.

First, the residual unit is expressed as:

$$y_l = h(x_l) + F(x_l, W_l)$$
$$x_{l+1} = f(y_l) \quad (1)$$

where x_l and x_{l+1} represent the l input and output of the th residual unit, respectively. F is the residual function, which represents the learned residual, and $h(x_l) = x_l$ represents the identity mapping, which f is the ReLU activation function. Based on the above formula, we obtain l the learning features from shallow $x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i)$to deep as: L

Using the chain rule, the gradient of the reverse process can be found:

$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_l} \sum_{i=l}^{L-1} F(x_i, W_i)\right) \quad (2)$$

of the loss function represented by L the first factor of the formula, $\frac{\partial loss}{\partial x_L}$ the first in the parentheses indicates that the short-circuit mechanism can propagate the gradient losslessly, while the other residual gradient needs to pass through the layer with weights, and the gradient is not directly passed. of. The residual gradient is not so coincidentally all -1, and even if it is small, the presence of 1 will not cause the gradient to disappear. So residual learning will be easier.

ResNet is divided into the following 5 different versions according to the number of layers.

TABLE I. FIVE VERSIONS OF RESNET

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

The network structure of ResNet-50 used to construct the driving fatigue detection model in this study is shown in Figure 4.

Moreover, ResNet uses two types of residual units, as shown in Figure 6. The left image corresponds to a shallow network, while the right image corresponds to a deep network. For short-circuit connections, when the input and output dimensions are the same, the input can be directly added to the output. But when the dimensions are inconsistent (corresponding to doubling the dimensions), this cannot be directly added. There are two strategies: (1) Use zero-padding to increase the dimension, in this case, you generally need to do a downsamp first, we can use pooling with stride=2, which will not increase the parameters; (2) Use a new mapping (projection shortcut), Generally, a 1x1 convolution is used,

which will increase the parameters and increase the amount of calculation. In addition to using identity mapping directly, short-circuit connections can of course use projection shortcuts.

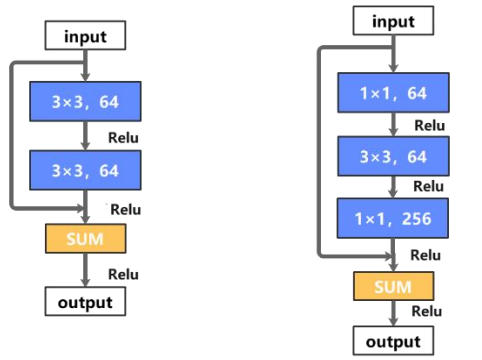


Fig. 4. Two types of residual units of ResNet.

Assuming that the inputs of both residual structures are 3 × 3 × 256 features, the total number of parameters of the left residual structure in Figure 4 is 3×3×256×256×2=1179648, and the total number of parameters on the right is 1×1×256 ×64+3×3×64×64+1×1×64×256=69632, therefore the residual structure on the right can greatly reduce the amount of calculation and parameters and speed up the training of the network.

To sum up, ResNet-50 reduces the amount of parameters calculated, solves the problem of gradient explosion or gradient disappearance through the residual structure, and has the advantages of fast training network speed and deep level. Based on this, this paper chooses to build a driving fatigue detection model through ResNet-50.

IV. Experiment Analysis and Results

A. Apparatus and Dataset

PyTorch is used to develop convolutional neural learning framework algorithms in this paper. Moreover, the operating system was Windows 11 and GeForce RTX 3070 as GPU.

The dataset used for training and testing in this study is the Distracted Driver Dataset [23] [24] from the Machine Intelligence Group at the American University in Cairo (MI-AUC).

Distracted Driver Dataset was collected by using a rear-facing camera installed on an ASUS ZenPhone (Model Z00UD). Input was collected in the form of a video and then cut into images of 1920 × 1080 pixel. The data owner emphasizes that all driving activities in the dataset were recorded while the participant was actually driving, not at a stationary location in a parking lot.

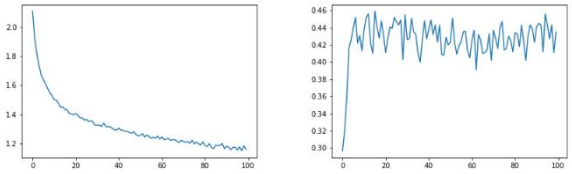
The study has 31 participants from seven different countries: Egypt (24), Germany (2), the United States (1), Canada (1), Uganda (1), Palestine (1), and Morocco (1). There were 22 males and 9 females in the study. Moreover, all the images were collected in 4 different cars: Proton Gen2 (26), Mitsubishi Lancer (2), Nissan Sunny (2), and KIA Carens (1). What's more, all 17,308 frames are divided into the following 10 categories: Drive Safe (3,686), Talk Passenger (2,570), Text Right (1,974), Drink (1,612), Talk Left (1,361), Text Left (1,301), Talk Right (1,223), Adjust Radio (1,220), Hair & Makeup (1,202), and Reach Behind (1,159).

B. Experiment and Result

On the basis of these categories, two sets of experiments were conducted:

Experiment 1: Training the data according to the ten categories in the Distracted Driver Dataset and test the learning effect.

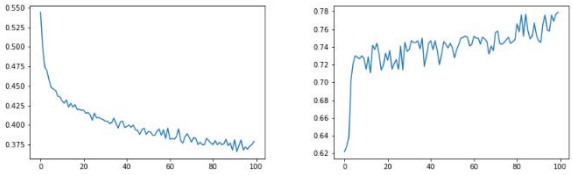
Experiment 2: The nine unsafe driving behaviors are uniformly summarized as unsafe driving states, and the two classifications are trained and evaluated.



(a) loss (b) accuracy

Fig. 5. Curves of loss and accuracy in Experiment 1.

In Experiment 1, the final values of loss and accuracy after training the neural network were 1.161 and 0.435, respectively, and the curves of loss and accuracy over time are shown in Figure 5.



(a) loss (b) accuracy

Fig. 6. Curves of loss and accuracy in Experiment 2.

Figure 6 illustrates the curves of loss and accuracy over time for Experiment 2. The final values of loss and accuracy for this experiment are 0.379 and 0.779, respectively.

It can be inferred from the figures that:

(1) The training effect is greater when all categories are classified strictly based on whether it is safe to drive or not.

(2) From the curve trend, Experiment 1 performs better in terms of loss convergence, but the fluctuation in accuracy does not increase toward the end of the test. Experiment 2 has a positive performance in loss convergence and an upward trend of accuracy, but the fluctuation is also larger. It proves that only two categories are more appropriate for the current experiment.

After the training, the validation was performed with the images that were not involved in the training, and the validation results of experiment 1 are shown in Figure 7. The

twelve images selected in the experiment include 3 safe driving state and 9 different categories of distracted driving.

Fig. 7. Prediction results in Experiment 1.

Fig. 8. Prediction results in Experiment 2.

The twelve images selected in Experiment 2 are the same as those in Experiment 1, and the classification results are presented in Figure 8. With only the twelve images we selected, the Probably for the output of Experiment 2 is above 0.7, which is significantly better than the value of Experiment 1.

In addition to the images with correct predictions in Figures 7 and 8, Figures 9 and 10 present the result of two groups in each of Experiment 1 and Experiment 2.

Fig. 9. Failed Prediction result in Experiment 1.

Figure 9 shows a safe driving image on the left, which is referred to as the "Text Right" state. Our hypothesis is that it may be because the driver put the telephone on his lap at this time, which affects the judgment of the neural network. The classification in the right figure should be "Talking to Passenger", but predicted as "Safe driving" here.

Fig. 10. Failed Prediction result in Experiment 2.

Figure 10 shows the data of the two failed prediction results in Experiment 2, both of which identified safe driving as unsafe driving, which may be related to the telephone on their lap or the hand off the steering wheel. From the above experimental results, it can be guessed that there may be some variability in the performance of specific actions of different people who concentrate on driving due to different driving habits.

V. Conclusion and Future Prospect

This paper proposes a method for classifying driver states using convolutional neural networks and validates it using the dataset Distracted Driver Dataset. The experimental results obtained after 100 iterations of training on the ResNet network indicate that a convolutional neural network can be useful in classifying driver states. It is evident that the performance of the model is unsatisfactory for classifying specific distraction categories of drivers, but the accuracy can reach 0.779 when only determining whether the driver is distracted or not. It is believed that the experimental results have further room for improvement.

- As we can see from the loss curve, the loss is convergent, so the model can perform better by increasing the number of training sessions.
- The input images could perhaps be pre-processed by histogram or morphology to make the environmental conditions more uniform and help the deep neural network to learn and recognize

The results of these experiments indicate that the current neural network application for driver distraction detection is not accurate, and we will suggest some research directions to pursue in the future.

- Despite the fact that ResNet-50 is not perfect, it does not prove that it does not have potential for improvement, much less that other neural networks will not be more efficient. In future research, other neural networks can be used for the task of driver distraction detection, or a special neural network can be proposed for this topic.
- The state of the driver captured in the cabin is dynamic, and it can be difficult to infer whether the driver is distracted by a small action at a particular moment, so future experiments may utilize the average classification of a video sequence to determine the state of the driver.
- Due to different driving habits, it is difficult to judge the driving status of different drivers by a unified standard. Therefore, in order to improve the accuracy of the detection, the relevant researchers can create a proprietary data set of driving states for different drivers, and perform targeted training based on this data.

REFERENCES

[1] O. I. d. C. d. OICA, "Automobiles.(2019b). Production Statistics, OICA," 2019.

[2] A. Fernández, R. Usamentiaga, J. L. Carús, and R. Casado, "Driver Distraction Using Visual-Based Sensors and Algorithms," *Sensors,* vol. 16, no. 11, pp. 1805, 2016.

[3] T. A. Ranney, W. R. Garrott, and M. J. Goodman, *NHTSA driver distraction research: Past, present, and future*, Citeseer, 2001.

[4] A. Guettas, S. Ayad, and O. Kazar, "Driver State Monitoring System: A Review," in Proceedings of the 4th International Conference on Big Data and Internet of Things, Rabat, Morocco, 2019, pp. Article 28.

[5] A. Halin, J. G. Verly, and M. Van Droogenbroeck, "Survey and Synthesis of State of the Art in Driver Monitoring," *Sensors,* vol. 21, no. 16, pp. 5558, 2021.

[6] J. Gonçalves, and K. Bengler, "Driver State Monitoring Systems–Transferable Knowledge Manual Driving to HAD," *Procedia Manufacturing,* vol. 3, pp. 3011-3016, 2015/01/01/, 2015.

[7] T. A. Ranney, "Driver distraction: A review of the current state-of-knowledge," 2008.

[8] J. Lee, M. Regan, and K. Young, *Defining Driver Distraction. Driver Distraction: Theory, Effects*, and Mitigation.

[9] A. Sonnleitner, M. S. Treder, M. Simon, S. Willmann, A. Ewald, A. Buchner, and M. Schrauf, "EEG alpha spindles and prolonged brake reaction times during auditory distraction in an on-road driving study," *Accident Analysis & Prevention,* vol. 62, pp. 110-118, 2014/01/01/, 2014.

[10] A. Braun, S. Frank, M. Majewski, and X. Wang, "CapSeat: capacitive proximity sensing for automotive activity recognition," in Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Nottingham, United Kingdom, 2015, pp. 225–232.

[11] M. A. Recarte, and L. M. Nunes, "Effects of verbal and spatial-imagery tasks on eye fixations while driving," *Journal of experimental psychology: Applied,* vol. 6, no. 1, pp. 31, 2000.

[12] M. A. Recarte, and L. M. Nunes, "Mental workload while driving: effects on visual search, discrimination, and decision making," *Journal of experimental psychology: Applied,* vol. 9, no. 2, pp. 119, 2003.

[13] R. van der Horst, "Occlusion as a measure for visual workload: an overview of TNO occlusion research in car driving," *Applied Ergonomics,* vol. 35, no. 3, pp. 189-196, 2004/05/01/, 2004.

[14] L. Chin-Teng, L. Hong-Zhang, C. Tzai-Wen, C. Chih-Feng, C. Yu-Chieh, L. Sheng-Fu, and K. Li-Wei, "Distraction-related EEG dynamics in virtual reality driving simulation." pp. 1088-1091.

[15] [K. Kircher, "Driver distraction: A review of the literature," 2007.

[16] P. Dhruv, and S. Naskar, "Image Classification Using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN): A Review." pp. 367-381.

[17] A. A. M. Al-Saffar, H. Tao, and M. A. Talab, "Review of deep convolution neural network in image classification." pp. 26-31.

[18] R. J. Hassan, and A. M. Abdulazeez, "Deep learning convolutional neural network for face recognition: A review," *International Journal of Science and Business,* vol. 5, no. 2, pp. 114-127, 2021.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." pp. 770-778.

[20] B. Alotaibi, and M. Alotaibi, "A hybrid deep ResNet and inception model for hyperspectral image classification," *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science,* vol. 88, no. 6, pp. 463-476, 2020.

[21] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognition,* vol. 90, pp. 119-133, 2019.

[22] A. Ajit, K. Acharya, and A. Samanta, "A review of convolutional neural networks." pp. 1-5.

[23] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," *arXiv preprint arXiv:1706.09498*, 2017.

[24] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, "Driver distraction identification with an ensemble of convolutional neural networks," *Journal of Advanced Transportation,* vol. 2019, 2019.