Enhancing Road Safety: Real-time Drowsiness Detection System

Ajay Singh[1], Subrata Sahana[2]
[1,2]*Department of Computer Science and Engineering, School of Engineering and Technology*
Sharda University, Greater Noida, UP, 201306, India

Abstract— This paper presents a novel real-time drowsiness detection system that aims to increase road safety by reducing the number of accidents caused by driver fatigue. The system's continuous monitoring of the driver's facial features, eye movements, and head posture is achieved by using an in-vehicle camera machine learning and computer vision. A CNN (Convolutional Neural Network) interprets the video frames and diagnoses various drowsy states such as prolonged eyelid closures and head tilts. Moreover, sensor information — such as steering patterns and braking — as well as vehicle speed — are also utilized to increase the detection accuracy. The system generates drowsy driver auditory and visual alerts instantaneously drowsy drivers to help make immediate actions. High experimental accuracy and responsiveness reveal that the proposed system can efficiently aid in road traffic accidents reduction, thus saving lives.

Keywords— Deep Learning; YOLOv8, Attention Detection, Multi-scale Feature Extraction; Swin Transformer

I. INTRODUCTION

Road safety is an important global issue with millions of lives being lost each year due to traffic accidents. The World Health Organization (WHO) reports that about 1.3 million people die in road crashes every year, in addition to many who become seriously injured. One of the main reasons for these accidents is driver fatigue which is distinguished by a significant decrease in alertness and slower reaction times. Oftentimes, drowsy drivers fail to recognize that they are no longer capable of operating a vehicle safely, and as a result, fatigue-related crashes happen without warning. These injuries are particularly bad because the drivers who are drowsy tend to not take precautionary measures before hitting other vehicles which causes more fatalities than other types of accidents. The solution to this challenge lies in creating new ways of detecting drowsiness in real time and subsequently controlling the situation.

A comprehensive drowsiness detection system covers a variety of fields such as machine learning, computer vision, human physiology, and automotive engineering. Such a system generally consists of a set of in-vehicle sensors and cameras that are used to observe and analyze the driver's physical as well as behavioural signs. Cameras, for instance, can take images of the facial expressions, eye movements, and head postures; sensors, in turn, can be used to check steering patterns, lane deviation, and vehicle speed. Sophisticated algorithms carry out this task using data streams to detect drowsiness such as deep eyelid closure, excessive yawning, or an erratic driving pattern. Hence, through the fusion of these technologies, the system can offer instantaneous feedback, thus, allowing drivers to correct any potential hazards before they arise.

The foundation of any drowsiness detection system is its ability to analyze data fastly and accurately. The key factor of machine learning in this development is teaching the system to learn from different data sets and strengthening its outcome as time goes by. Convolutional Neural Networks (CNNs), as an illustration, are very good at analyzing visual data, like images or video streams of the driver's face. These networks can catch small changes in facial features such as drooping eyelids or lowering blink rates which may signify fatigue. Likewise, Long Short-Term Memory (LSTM) networks are also useful in examining time-series data such as steering inputs or eye movement patterns to diagnose drowsiness.

Besides machine learning, the drowsiness detection system must also take into consideration the design practicalities for implementation. For example, there can be substantial changes in indoor lighting conditions inside the car — this will affect the validity of visual data captured by cameras. Likewise, drivers may wear accessories such as sunglasses or face masks, which can hide facial features. The need to solve these problems will be met with proper preprocessing techniques and the usage of multi-modal data to improve system reliability. In addition,

the system must function in real-time, thus alerts must be sent out with the least possible delay. For this purpose, algorithms that are fast and optimized hardware that can do complicated calculations without reducing performance will be needed.

Besides this, the other important part is drowsiness detection should be integrated into the whole vehicle safety system ecosystem. Vehicles nowadays come with many sensors as well as connectivity features like GPS, accelerators, and vehicle-to-vehicle communication systems. The drowsiness detection system can use these technologies to get a deeper perspective on driver behaviour as well as the surrounding conditions. For example, the system might utilize the GPS data for the identification of a long monotonous road which is likely to increase fatigue. Furthermore, the vehicle-to-vehicle communication system might enable the system to send drowsiness alerts to other cars thereby improving the road safety level.

A real-time drowsiness detection system can be generalized in terms of its social and economic implications. On a societal level, by cutting the rate of weariness-induced accidents down to a minimum, thousands of lives and numerous injuries might be saved every year. From an economic perspective, the reduction of road traffic injuries, consequently, decreasing medical costs, property damage, and lost productivity, can result in considerable savings. Moreover, the widespread application of such systems can build the foundation of public trust in automotive technology facilitating its use for more advanced safety features and eventually, fully autonomous vehicles.

Even though the introduction of systems that can detect drowsiness has a beneficial impact, a lot of people choose not to adopt these systems. The main obstacle to implementation is cost, especially for budget vehicles. Although high-end vehicles are often equipped with standard high-level safety features, low-cost models require integrating such systems in a cost-effective manner. Furthermore, the usage of cameras and sensors inside vehicles for data collection makes the privacy issue a very serious one. Guidelines on data storage, usage, and protection must be clear to the public to ensure the technologies are accepted.

This research project is all about the development and carrying out of a drowsiness detection system in real-time, and it mainly focuses on the technical parts, practical difficulties, and possible use in increasing road safety. The study will analyze existing research and propose solutions with the aim of aiding the efforts that are already ongoing to make roads safer for everyone. The following parts of the dissertation will deal with the core technologies, system architecture, experimental results, and future directions of this promising field of research.

In short, real-time drowsiness detection is one of the most important activities that can be put in place to ensure road safety, hence reducing the number of traffic accidents globally. This is achievable through the harnessing of machine learning technology, computer vision, and automotive engineering to come up with such systems that not only detect drowsiness with a high level of accuracy but also work in a seamless manner with most modern vehicles. As this technology advances, it has the potential to redefine driving, thus roads will be safer and lives will be saved around the globe.

II. RELATED WORK

A. Vision-based Drowsiness Detection Systems

There have been numerous studies about the application of computer vision to the problem of drowsiness detection. Scientists, for example, created systems that utilize facial feature analysis such as eye blink frequency, eyelid closure duration, and yawning detection to determine a level of fatigue. Convolutional Neural Networks (CNNs) are used to process pictures or movie clips taken by in-vehicle cameras and the results achieved are in high accuracy of controlled environments. Nevertheless, these systems are not performing so well in other dynamic conditions, for instance, in the case of changing lighting or the use of sunglasses, which is their main disadvantage in real-life situations.

B. Multi-modal Sensor Fusion Approaches

One of the perspectives is to fuse several data streams like visual and non-visual inputs to improve drowsiness detection. There have been some studies that included accompanying biofeedback sensing such as heart rate and EEG (Electroencephalogram) data with facial feature detection (FFD) using face sensing to analyze steering behaviour, and lane deviation patterns. These systems based on more than one input reduce the dependence on one input, strengthening the argument that environmental factors or driver variability may affect the result. However, the increased complexity and cost related

to the implementation of such systems have been the reasons for the limited distribution of these technologies.

C. Machine Learning and Time-Series Analysis for Driver Behavior

Driving behaviour time-series data have been examined through machine learning models like Long Short-Term Memory (LSTM) networks and Support Vector Machine (SVMs). These systems detect gradual patterns over time, such as steering wheel movement or gradual changes in speed to identify drowsiness trends. Although they are good at finding long-term indicators of fatigue, these methods also have challenges in providing instantaneous feedback, which is very important in real-time applications. The new hybrid models that combine CNNs and LSTMs have been successful in bringing this together through the fusion of spatial and temporal data for real-time detection.

III. RELATED IMPROVEMENT WORK

Backbone Network Enhancement, the uses of this algorithm democratically detect objects, using the CSPDarknet framework which is a convolution neural network as a backbone network. Other modules to enhance the detection of small-sized objects are stacked atop the SPPF in the YOLOv5 backbone.

To be specific, after the SPPF stage in the feature extraction, a new module and three convolution operations are introduced. The input feature map goes into two branches: one with convolving and modular processes and the other with only convolutions. Outputs from both branches are concatenated and passed through another convolution operation. The module consists of six different convolution operations, thus enabling feature extraction for various given fields. By stacking convolution layers, the receptive field is broadened, making it possible to extract features across multiple scales. The outputs from these receptive fields are concatenated to enhance feature perception, particularly for detecting small objects.

The Swin Transformer was introduced by Microsoft in March 2021. This general-purpose backbone for computer vision, specifically designed for image segmentation and detection, has been very successful.

In the Swin transformer, the two concepts of patched sliding window operations and hierarchical designs are employed to address the high computational cost engendered by larger visual entities and/or high-resolution images. It has three major components: Patch Embedding, Swin Transformer Block, and Patch Merging. Together, those components mitigate computational overhead while convincingly preserving accuracy. The Swin Transformer tantamounts to success on some of the complex visual tasks.

IV. METHODOLOGY

In this section, we discuss the proposed methodology for the driver drowsiness system which contains two phases namely, face extraction and drowsiness detection. We first use facial landmarks to extract faces and then using YOLOv5 detect drowsiness. We use appropriate data augmentation techniques so that we increase the amount of data and thereby enhance the performance. The schematic representation of the proposed system is depicted in Figure. 1.

A. Data Augmentation:

In this work, data augmentation is only applied to the custom dataset and not on the benchmark UTA dataset. Appropriate data augmentation techniques are applied on the custom dataset, which increases the size and consequently improves the model performance. Data augmentation is represented as A: X→ X', where X' is the set of augmented images. Various transformations include rotation, scaling, and flipping that obtain the augmented dataset X' = {X1′, X2′, X3′…,Xn′}.

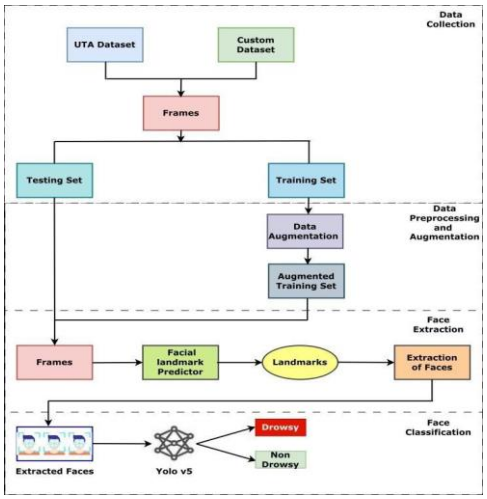

Fig. 1. Illustration of proposed drowsiness detection framework with data collection, data augmentation, facial landmark identification, and YOLOv5- based drowsiness detection modules.

B. Face extraction using facial landmark

Face alignment refers to the process of locating important facial landmarks, such as eyes, nose and mouth, and adjusting the image to put those landmarks in an aligned way with each other. Common computer vision techniques used in face alignment are facial landmark detection and affine transformations.

Facial landmark detection involves processing the algorithms for the identification and location of these key facial features, which include the corners of the eyes and nose tip, and the corners of the mouth. These are further used as landmarks for aligning the images. An affine transformation is a set of mathematical operations that allows one to rotate, scale, or even translate an image. The facial landmarks identified can be used to transform the image so that the landmarks fall within a standardised position relative to each other. For this example, the eyes would be positioned horizontally, and the mouth vertically aligned. The aligned image can then be used in applications such as face recognition, facial expression analysis, and virtual try-ons.

Given a set of images X = {X1, X2, X3…., Xn}, where Xi represents the i*th* image containing the driver's face, and the corresponding labels Y = {Y1, Y2, Y3…., Yn} where Yi is the binary label of the image i*th* indicating a drowsy (1) or non-drowsy (0) state. For each image, Xi, the custom face extraction model is put on to identify facial landmarks, represented as Li = {l1, l2, l3, ., ln}, where li represents the lm facial landmark point. This results in the output of the face extraction phase as a set of facial landmarks L = {L1, L2, L3, ., Ln}, corresponding to each image Xi in the dataset. Facial landmarks are used here to detect faces. We incorporated two types of landmark predictors: (i) a 68-point landmark predictor as in Figure 2 and (ii) a 5-point landmark predictor.

The performance of face landmarking was assessed using the normalised root mean square error (NRMSE), which was normalised concerning the inter-ocular distance (IOD). The IOD is defined as the distance between the centres of the two eyes. By normalising the errors in landmark localisation with the IOD, the evaluation of performance becomes impartial to the size of the face or the zoom factor of the camera.

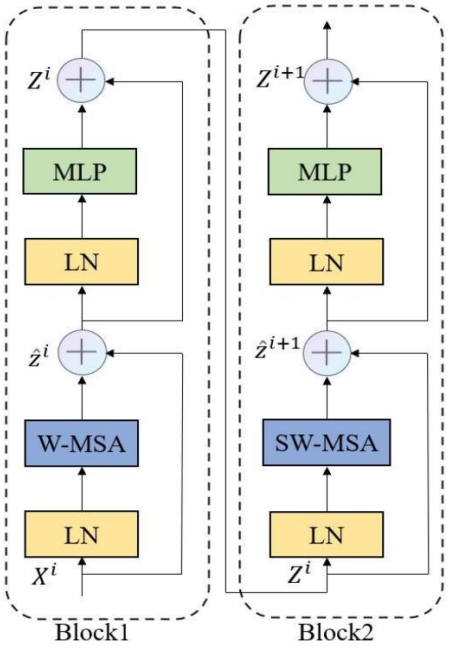

Fig. 2. Swin Transformer Block Architecture

The normalized distance δ is computed as the Euclidean distance d(. , .) between the ground-truth landmark coordinates (x, y) and the predicted landmark coordinates(˜x̃, ỹ), normalized by the IOD.

(Red dots represent ground truth landmarks, and green dots represent predicted landmarks).

V. EXPERIMENTAL RESULT

A. Introduction to Experimental Dataset In order to evaluate whether the accuracy of the model for detecting human attention has improved after the improvements, the VOC open-source dataset was used in this experiment. About 3500 images of smoking, drinking, and mobile phone use were selected and labelled, and there were four categories of labels: face, smoke, drink, and phone. The dataset was divided into training and validation sets in a 9:1 ratio. During the preprocessing stage, Mosaic data augmentation was used to scale and z̄stitch together any four images, which enriched the dataset to some extent and strengthened the robustness of the network. The experimental environment configuration and evaluation metrics are introduced below, followed by a comparison with the model before improvement through ablation experiments and presentation of the experimental results. 5.2. Experimental Environment Configuration The operating system of this experiment is the Win10 Chinese version. The main hardware component of the experiment is an Intel (R) Core (TM) i7-10875H CPU; 16G memory; and an NVIDIA GeForce RTX2060 GPU. The deep learning framework

adopted is PyTorch. When training, the epoch was set to 150, the batch size was set to 16, and the optimizer selected was SGD.

VI. CONCLUSIONS

There is this new approach to identifying when drivers get distracted at the wheel, summed up into two main areas: distracted abnormal behaviour detection. The method effectively addresses the challenges associated with both these types of distracted driving. The three common types of distracted behaviours observed were especially highlighted: smoking, drinking water, and using a mobile phone.

To identify drowsiness, the proposed approach depends on the aspect ratio computation between the width and height of the eyes and mouth for a driver, and it can effectively detect the signal of sleepiness. For distracted behaviour identification, the research proposal is an enhanced version of YOLOv5. The approach is updated by adding more convolution operations within the backbone network to generate a diverse receptive field and to extract more useful features from the system.

Additionally, the Swin Transformer module is applied to replace one Bottleneck module in the C3 module of the network of feature fusion. Such substitution heightens the model's sense of global information. Besides, it further strengthens the connection patterns of the feature fusion network and consequently enhances the overall model's feature fusion capability. These advantages significantly improve the ability of the model the detect small objects such as cigarettes and water bottles, thereby making contributions to better distracted behaviour detection.

Experimental results confirm that the proposed method has an improved mean Average Precision (mAP) in comparison with the original YOLOv5 model. The system also shows that it is related to and benchmarked with the newer YOLOv7 model, where the size of the model is relatively smaller; however, detection accuracy for the proposed method is higher. However, it can be observed that the detection speed of the proposed model is slower compared to the original YOLOv5. Such a limitation will be the area of focus in future work aimed at enhancing the real-time performance of the network model.

It is worth noting that the three types of distracted behaviours encompassed here are in no way exhaustive of the types of distracted behaviours toward which traffic safety may be compromised. Other distracting behaviours related or that might put road safety at risk have not been considered within these study boundaries; hence, this study cannot claim to analyze comprehensively all distracting behaviors. Further research will therefore go further to include other distracting behaviors, improving the scope by widening the application and effectiveness of the model in real-life contexts.

In summary, the current study provides a broad solution to driver fatigue and distracted behaviours through the application of high-performance deep learning techniques, where an improved feature extraction scheme combined with a global information-aware module helps raise the overall detection accuracy for small objects by large margins. While detection speed and the challenges mentioned above would have continued to be an issue, it should be tolerable to assume that those who continue to further improve such technology would have addressed those weaknesses by opening a very effective system of real-time monitoring of drivers that could lead to safer roads for the world.

REFERENCES

[1]. Abdullah, M., Agal, A., Alharthi, M., & AL Rashidi, M. (2018). Retracted: Arabic handwriting recognition using neural network classifier. Journal of Fundamental and Applied Sciences, 10(4S), 265-270.

[2]. Abe, S. (2010). Support Vector Machines for Pattern Classification. Berlin, Germany: Springer Science & Business Media.

[3]. Aggarwal, C. C. (2018). Neural Networks and Deep Learning: A Textbook. Basingstoke, England: Springer.

[4]. Balas, V. E., Roy, S. S. Sharma, D.& Samui, P. (2019). Handbook of Deep Learning Applications. Basingstoke, England: Springer.

[5]. Boukharouba, A., & Bennia, A. (2017). Novel feature extraction technique for the recognition of handwritten digits. Applied Computing and Informatics, 13(1), 19-26.
Doi:10.1016/j.aci.2015.05.001

[6]. Buckland, M. K. (2006). Emanuel Goldberg and His Knowledge Machine: Information, Invention, and Political Forces. Santa Barbara, CA: Greenwood Publishing Group.

[7]. Chandio, A. A., Leghari, M., Hakro, D., AWAN, S., & Jalbani, A. H. (2016). A Novel Approach for Online Sindhi Handwritten Word Recognition using Neural Network. Sindh University Research Journal- SURJ (Science Series), 48(1).
[8]. Chen, L., Wang, S., Fan, W., Sun, J., & Naoi, S. (2015). Beyond human recognition: A CNN-based framework for handwritten character recognition. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 695-699. doi:10.1109/acpr.2015.7486592.
[9]. Ding, S., Zhao, H., Zhang, Y., Xu, X., & Nie, R. (2015). Extreme learning machine: algorithm, theory and applications. Artificial Intelligence Review, 44(1), 103- 115.
[10]. Dwivedi, U., Rajput, P. Sharma, M. K., & Noida, G. (2017). Cursive Handwriting Recognition System Using Feature Extraction and Artificial Neural Network. Int. Res. J. Eng. Techno, 4(03), 2202-2206.
[11]. El-Sway, A., Loey, M., & El-Bakry, H. (2017). Arabic handwritten character recognition using convolutional neural network. WSEASTransactions on Computer Research, 5, 11-19.
[12]. Ganapathy, V., & Liew, K. L. (2008). Handwritten Character Recognition Using Multiscale Neural Network Training Technique. World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering, 2(3), 638-643.
[13]. Grady, J. O. (2010). System Requirements Analysis. Amsterdam, Netherlands: Elsevier.
[14]. Gribaudo, M. (2013). Theory and Application of Multi- FormalismModeling. Hershey, PA: IGI Global.
[15]. Gunawan, T. S., Noor, A. F. R. M., & Kartiwi, M. (2018). Development of English handwritten recognition using deep neural network. Indonesian Journal of Electrical Engineering and Computer Science, 10(2), 562-568.
[16]. Hertz, J. A. (2018).Introduction to the theory of neural computation. Boca Raton: CRC Press.Hamid, N. A., & Sjarif, N. N. A. (2017). Handwritten recognition usings, KNN and neural network.arXiv preprint arXiv:1702.00723.
[17]. Kumar, P., Saini, R., Roy, P. P. & Pal, U. (2018). A lexicon-free approach for 3D handwriting recognition using classifier combination. Pattern Recognition Letters, 103, 1-7.
[18]. Kirova, V., Manolopoulos, Y., Hammer, B., Iliadis, L., &Kalogiannis, I. (2018). Artificial Neural Networks and Machine Learning – ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings. Basingstoke, England: Springer.
[19]. Larasati, R., & KeungLam,H. (2017, November). Handwritten digits recognition using ensemble neural networks and ensemble decision tree. In 2017 International Conference on Smart Cities, Automation &Intelligent Computing Systems (ICON- SONICS) (pp. 99-104). IEEE.
[20]. digits recognition using ensemble neural networks and ensemble decision tree. In 2017 International Conference on Smart Cities, Automation &Intelligent Computing Systems (ICON- SONICS) (pp. 99-104). IEEE.