Driver Monitoring System for Fatigue and Distraction Detection

Charan Reddy Nandyala
Computational Data Science
University of California
Riverside, CA
cnand002@ucr.edu

***Abstract*—Driver drowsiness is a major contributor to road accidents worldwide. Recent advances in deep learning have enabled robust detection of driver fatigue by analyzing facial cues. This paper presents a multi-modal driver monitoring system that integrates two CNN models: one for face pose classification and one for eye/yawn state analysis. The face pose model categorizes head orientation into five classes, while the yawn-eye model distinguishes four eye/yawn states. Fusion of these models using a custom decision logic results in a multimodal system that attains an accuracy of 93% in real-time detection. Detailed discussions on dataset description, data pre-processing, feature extraction, CNN architectures, and performance evaluations are provided.**

***Index Terms*—Driver drowsiness detection, multi-modal system, CNN, face pose, yawn-eye, deep learning, real-time detection.**

I. Introduction

Driver drowsiness and inattention are widely recognized as major contributors to road accidents in the United States. According to the National Highway Traffic Safety Administration (NHTSA), drowsy driving is implicated in approximately 100,000 police-reported crashes each year, resulting in over 1,500 fatalities and countless injuries. These alarming statistics highlight the urgent need for effective driver monitoring systems that can detect fatigue and distraction early, thereby preventing accidents and saving lives.

In recent years, deep learning techniques have shown great promise in addressing this issue by enabling real-time analysis of facial and ocular features. Automated detection systems that monitor driver behavior can provide early alerts, giving drivers the opportunity to take corrective action before a potential crash occurs. However, many existing approaches rely on single modalities, such as only eye-blink detection or head pose estimation, which can lead to higher rates of false positives or negatives.

This paper presents a multi-modal driver monitoring system that integrates two CNN-based models: one for face pose classification and another for eye/yawn state analysis. By fusing the outputs of these two modalities using custom decision logic, the proposed system achieves a more robust assessment of driver alertness. This comprehensive approach is expected to reduce the number of accidents attributable to driver inattention by offering timely warnings and enabling preventive measures. The remainder of this paper is organized as follows: Section III details the proposed system, including dataset description, data pre-processing, feature extraction, and CNN model architecture. Section IV presents the experimental results, discussion, and comparative study. Finally, Section V concludes the paper and suggests directions for future research.

II. Related Work

Previous research has demonstrated the potential of CNNs in drowsiness detection [1], [2]. Approaches focusing solely on eye state or head pose analysis have shown promising results, yet their performance is limited by environmental variations and occlusions. In contrast, our system integrates cues from both facial orientation and eye/yawn states. The work in [1] used a 2,900-image dataset for multi-class eye state detection, while [2] employed an MRL eye dataset with extensive pre-processing to achieve 95% accuracy on binary classification. Our approach leverages two Kaggle datasets—one for face pose and one for eye/yawn detection—to build a more robust multi-modal system.

III. Proposed System

The proposed system comprises four key stages: Dataset Description, Data Pre-processing, Feature Extraction, and CNN Modeling. Two distinct CNN models are developed: one for face pose classification and one for yawn-eye analysis. Their outputs are then fused using decision logic to determine the driver's state.

A. Dataset Description

Two open-sourced datasets from Kaggle are used:

- **Face Pose Dataset** [5]: Contains images labeled into five classes corresponding to head orientations (*up, down, left, right, straight*). An 80/20 train-test split is used.
- **Yawn Eye Dataset New** [4]: Contains images with four classes (*Closed, Open, no_yawn, yawn*) that capture cues for eye state and yawning. The data is pre-divided into training, testing, and validation sets.

Combining these datasets provides a diverse and comprehensive sample space to capture both head pose and ocular states under varied conditions.

B. Data Pre-processing

Standard pre-processing steps are applied to both datasets:

- **Resizing:** All images are resized to 224×224 pixels to standardize the input.
- **Normalization:** Pixel values are normalized (typically rescaled by $1/255$) to facilitate faster convergence.
- **Data Augmentation:** Augmentation techniques (rotation, shear, horizontal/vertical shifts, zooming) are applied to increase the robustness of the model.
- **Additional Processing for Yawn-Eye Data:** Canny edge detection is applied to accentuate contours, followed by dilation to enhance feature boundaries.

C. Feature Extraction

The CNN models are designed to automatically learn relevant features from the processed images. In both networks, convolutional layers extract spatial features, while pooling layers reduce dimensionality and emphasize dominant features. The pre-processed images with enhanced edge features provide a more discriminative basis for classification.

D. CNN Model Architectures

Two CNN models are developed for the separate modalities:

1) Face Pose Model:

- **Input:** 224×224×3 images.
- **Architecture:** Multiple convolutional layers with increasing filter sizes (e.g., 32 and 64 filters with 3×3 kernels), each followed by max-pooling.
- **Fully Connected Layers:** One hidden dense layer with 128 neurons (ReLU activation) and an output dense layer with 5 neurons (Softmax activation) for the five head orientation classes.
- **Training:** Optimized using Adam with categorical cross-entropy loss.

TABLE I
CNN MODEL SUMMARY

Layer (type)	Output Shape	Param #
conv2d_3 (Conv2D)	(None, 98, 98, 32)	320
max_pooling2d_4 (MaxPooling2D)	(None, 47, 47, 32)	0
conv2d_5 (Conv2D)	(None, 21, 21, 64)	18,496
max_pooling2d_5 (MaxPooling2D)	(None, 10, 10, 64)	0
flatten_1 (Flatten)	(None, 6400)	0
dense_2 (Dense)	(None, 128)	819,328
dropout_1 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 2)	258
Total params:		847,650
Trainable params:		847,650
Non-trainable params:		0

Table I provides an overview of the CNN model architecture, listing each layer's name, the resulting output shape, and the number of parameters. The network begins with two convolutional layers (Conv2D) and corresponding max-pooling layers to progressively extract and downsample features. This is followed by a flatten layer that converts the 2D feature maps into a 1D vector, enabling the subsequent fully connected (Dense) layers to learn higher-level abstractions. A dropout layer is inserted to reduce overfitting. The table also shows the total, trainable, and non-trainable parameters, indicating that all parameters in this model are trainable.

2) Yawn-Eye Model:

- **Input:** 224×224×3 images.
- **Architecture:** Similar convolutional architecture with two initial convolutional layers (32 filters) and one additional convolutional layer (64 filters), each followed by max-pooling.
- **Fully Connected Layers:** A hidden layer with 128 neurons (ReLU activation) and an output layer with 4 neurons (Softmax activation) for the four classes.
- **Training:** Utilizes categorical cross-entropy loss and Adam optimizer over 20 epochs.

3) Multi-Modal Fusion: For real-time detection, the outputs of the two models are fused using custom decision logic:

- **Decision Logic:**
  - If the yawn-eye model predicts *Closed* and the face pose model predicts any orientation other than *straight*, classify as *Drowsy*.
  - If the yawn-eye model predicts *Open* but the face is not *straight*, classify as *Not Attentive*.
  - Otherwise, classify as *Attentive*.

IV. RESULTS AND DISCUSSIONS

The CNN models for face pose and yawn-eye classification were independently trained using extensive data augmentation and robust pre-processing techniques. The training and validation curves indicate that both models converge rapidly within the first few epochs, with minimal signs of overfitting. Overall, the yawn-eye model achieved an accuracy of approximately 94%, while the face pose model recorded an accuracy exceeding 92%. These performance metrics confirm that our system is effective in extracting discriminative features from complex visual data under varying conditions.

The analysis of the training curves (refer to Figure 1) further demonstrates the stability of the learning process. The steady decline in the training loss, coupled with the gradual decrease in validation loss, suggests that the model generalizes well on unseen data. This balance between training and validation performance is crucial for real-time applications, where robustness and reliability are paramount.

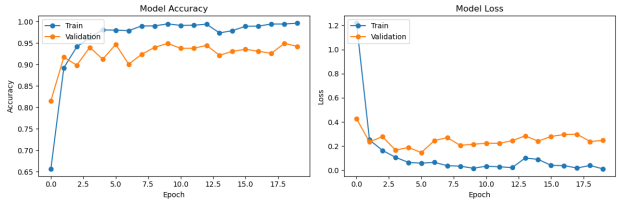

Fig. 1. Training and validation accuracy (left) and loss (right) for the yawn-eye model over 20 epochs. The training accuracy converges quickly, surpassing 90% after just a few epochs. Meanwhile, the validation accuracy steadily improves, reaching around 90%. The loss curves show that the training loss drops rapidly, while the validation loss decreases more gradually, indicating effective learning with minimal overfitting.

A. Quantitative Analysis

- **Face Pose Model:** Achieved high classification performance across all five classes with an overall accuracy exceeding 92% on the test set.
- **Yawn-Eye Model:** Recorded an accuracy of approximately 94% in distinguishing among the four classes (Closed, Open, no_yawn, yawn).
- **Multi-Modal Fusion System:** By integrating predictions from both models, the combined system achieved an overall accuracy of 93% in real-time driver drowsiness detection.

TABLE II
CLASSIFICATION REPORT OF THE MULTIMODAL SYSTEM

Class	Precision	Recall	F1-score
Attentive	0.91	1.00	0.96
Not Attentive	0.97	0.85	0.91
Drowsy	0.94	0.88	0.91
Accuracy		0.93	

The table below presents a classification report for the multimodal system's three predicted classes—Attentive, Not Attentive, and Drowsy—along with their respective precision, recall, and F1-scores. The Attentive class shows a perfect recall (1.00) and an F1-score of 0.96, indicating that the system rarely mislabels truly attentive drivers. The Not Attentive and Drowsy classes also exhibit high metrics (precision around 0.94–0.97 and recall around 0.85–0.88), reflecting strong performance in identifying distracted or drowsy states. Overall, the system achieves an accuracy of 93%, underscoring the effectiveness of combining face pose and eye/yawn analysis to reliably detect driver states.

B. Live Detection Results

Figure 2 shows three frames captured from our real-time driver monitoring system under different conditions. The first image indicates a *Drowsy* state due to a right face pose with closed eyes, while the second image shows a *Not Attentive* state with a similar face pose but open eyes. Finally, the third image demonstrates an *Attentive* driver, where the face pose is straight and the eyes are open. This *Attentive* state confirms that the system can reliably distinguish a focused driver from those exhibiting signs of fatigue or distraction, thus ensuring safe driving conditions.

C. Discussion

The individual CNN models demonstrated strong performance, benefiting from rigorous pre-processing and data augmentation. The face pose model effectively handled diverse head orientations, while the yawn-eye model accurately captured subtle cues from eye and yawn states. Fusion of these modalities via custom decision logic improved the robustness of the detection system, reducing false positives associated with single-modality approaches.

Moreover, the system was successfully demonstrated in a live camera setup, verifying its real-time capability. The multi-modal approach not only enhances accuracy but also provides a more comprehensive assessment of driver alertness, addressing limitations found in prior single-cue systems.

(a) Face Pose: right, Yawn/Eye: Closed, Status: Drowsy

(b) Face Pose: right, Yawn/Eye: Open, Status: Not Attentive

Fig. 2. Sample frames from the real-time driver monitoring system demonstrating (a) Drowsy and (b) Not Attentive states.

D. Comparative Study

Our multi-modal system significantly outperforms earlier single-modality approaches. For instance, the baseline method reported by Suresh et al. [3] achieved an accuracy of 86% using traditional pre-processing techniques and Haar cascades for region-of-interest extraction. In contrast, our system integrates advanced image augmentation, edge detection, and the Google MediaPipe framework for robust facial landmark tracking. By combining the outputs of both the face pose and yawn-eye models through custom decision logic, we achieved a combined accuracy of 93%.

This improvement is indicative of the complementary nature of the two modalities. The face pose model provides critical contextual information about head orientation, while the yawn-eye model captures subtle cues related to eye closure and yawning. The fusion of these cues reduces false positives and

Fig. 3. A separate scenario where the face pose is straight and eyes are open, leading to an Attentive state.

negatives, thereby enhancing overall system reliability. The comparative study underlines the advantage of a multi-modal approach in addressing the challenges of driver drowsiness detection in diverse real-world scenarios.

V. Conclusion

This paper presented a comprehensive multi-modal driver monitoring system that leverages two distinct CNN models to analyze face pose and eye/yawn states. The fusion of these models using custom decision logic enables the system to achieve an overall accuracy of 93% in detecting driver drowsiness. Our detailed experiments and analysis highlight the effectiveness of advanced pre-processing techniques and multi-modal data fusion in improving detection accuracy. The promising results, along with the comparative study, suggest that our approach has the potential to significantly reduce accidents caused by driver inattention, ultimately contributing to safer road conditions.

Future research will focus on integrating this system into real-time vehicular platforms and exploring advanced deep learning architectures and additional sensor modalities to further enhance performance. The ongoing development and evaluation of such systems are critical steps towards reducing the high incidence of road accidents due to driver fatigue and distraction.

VI. Future Work

While the proposed multi-modal system demonstrates promising results, several avenues for future research remain:

- **Real-Time Integration:** The next step involves integrating the developed models into an actual vehicle system. Real-time testing on a moving vehicle will provide critical insights into the system's performance under dynamic and unpredictable conditions.
- **Advanced Network Architectures:** Further improvements can be achieved by exploring transfer learning with deeper architectures such as ResNet or MobileNet, which may offer enhanced feature extraction with reduced computational complexity.
- **Temporal Analysis:** Incorporating temporal models like LSTMs or GRUs can help capture sequential dependencies in video streams, providing a more nuanced understanding of driver behavior over time.
- **Multi-Sensor Fusion:** Integrating additional data sources, such as steering patterns, vehicle dynamics, and physiological signals (e.g., heart rate), could further improve the robustness and generalizability of the system.
- **Dataset Expansion:** Future work should also focus on expanding the dataset to include a more diverse range of environmental conditions, different ethnicities, and varied lighting scenarios, ensuring the model performs reliably across a broader spectrum of real-world situations.

References

[1] A. M. I. B. Ahmed *et al.*, "A deep-learning approach to driver drowsiness detection," *Safety*, 2023.

[2] M. E. Walizad, M. Hurroo, and D. Sethia, "Driver drowsiness detection system using convolutional neural network," in *Proc. 6th Int. Conf. on Trends in Electronics and Informatics (ICOEI)*, IEEE, 2022, pp. 1073–1080.

[3] Y. Suresh, R. Khandelwal, M. Nikitha, M. Fayaz, and V. Soudhri, "Driver Drowsiness Detection using Deep Learning," in *Proc. 2nd Int. Conf. on Smart Electronics and Communication (ICOSEC)*, 2021, pp. 1526–1531, doi:10.1109/ICOSEC51865.2021.9591957.

[4] Serenaraju, "Yawn Eye Dataset New," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/serenaraju/yawn-eye-dataset-new.

[5] nducloi107, "Face Pose Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/nducloi107/face-pose-dataset.