IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

CHAUFFEUR BEHAVIOR RECOGNITION BASED ON THE CONVOLUTION NEURAL SYSTEM

B.L Jayakumar, Nimishambha Bharani S, Pushpa Yadav, Raksha. M

S.E.A COLLEGE ON ENGINEERING AND TECHNOLOGY

EKTA Nagar, Basavanapura, Vigonagar, Krishnarajapura, Bangalore,560036

Abstract:

Chauffeur behavior recognition is extensively used to lessen the threat of traffic accidents. maximum of the previous techniques for monitoring chauffeur's behavior is based on computer vision techniques. these strategies have the potential for invasion of privacy and spoofing. this text gives a brand new however effective deep studying approach for reading driving behavior using cues inclusive of facial expressions used to apprehend five forms of using: aggressive, distracted, drowsy, and drinking distraction.to make use of a successful deep neural network from snapshots, we study convolutional neural networks (CNNs) from photographs built from manipulated alerts based totally on the circular plot technique. As a result of the test, it changed into confirmed that the proposed technique can correctly detect the chauffeur's behavior.

Introduction:

Road accidents cause around 1.35 million fatalities and up to 50 million injuries yearly globally. Traffic-related mortality and injury cost the global economy around 518 bill

Although road quality, weather conditions, and vehicle performance all play a crucial role in accidents, human behavioral factors significantly impact the vast majority of accidents Studies show that in 95% of accidents, the human factor is critical and driver behavior is recognized as the most important factor. Specifically, in the United States, according to a 2016 study, it was found that only 7.27 million car accidents occurred, resulting in 37,914 deaths and 2.17 million injuries in which human factors had the most effect (about 94%). Many platforms can assist drivers in making safe trips and decrease road fatalities. One of these platforms is the Advanced Driving Assistance System (ADAS). Rather than. monitoring driver driving style, ADAS focuses on assisting drivers. These systems could protect many lives after accidents but could not prevent an accident from taking place. Due to the expensive price, ADAS implementation has been limited. Managers in freight management companies utilized real-time, continuous, and automated driver behavior profiling to institutionalize campaigns to improve drivers' scores, decrease the accident rate, increase the resource-based economy, and extend vehicle lifetime warranties. Moreover, to prevent accidents, we could notify drivers of aggressive driving events in real time. For example, a smartphone app could warn when the driver makes an aggressive U-turn. Considering all of the aforementioned aspects, driver behavior should be regarded as one of the most vital areas for improving road safety. In addition, driver behavior could impact the decrease in fuel consumption and greenhouse gas emissions, of green transportation systems Researchers, used different solutions to evaluate their driving event detection and behavior models. Some of them used the Driver Behavior Questionnaire to investigate the effect of safety skills on driver behaviors . These questionnaires are based on drivers' self-statements about their previous driving behaviors and experiences. However, they did not give insights into real-time driving events. In order to cover this limitation, in-vehicle black box systems were introduced to capture sensor data and driver behaviors simultaneously. Richard et al. used SHRP2 in-vehicle sensor data to examine real-time driver speeding behavior which primarily comprised video images of the front and rear windshields, rate of acceleration, and some sensor data. Nevertheless, SHRP2 had its limitations, such as Gyroscope and Linear Accelerator sensors were not used in this dataset, and there was no real-time labeling mechanism for driving events and driver behavior(except for reporting emergencies). We also analyzed SHRP2 real-time data to detect Right Turn driving events Considering such events and previous inventions we have come up with a new model. In this model, we are going to detect both driver and the person sitting next to the driver With the help of the camera the detection is done, and it displays the result on the screen for example if the driver is looking towards the right it detects and display's on the screen saying that "you are looking towards the right" similarly all the actions of both the people sitting in the front seat will be detected and displayed on the screen.

The most attention is given to the driver. In the previous inventions, the detectors were wearable devices that were very distracting to the driver and led to many accidents and they detected only behavior of the driverThe proposed network architecture consists of the feature extractor and classifier modules. The feature extractor module is designed based on the stem module, adaptive connections (AC), and a CBAM to extract the feature maps. At the final feature map, the classifier module applies a GAP and a softmax function to calculate the probability of ten driver behaviors and then classifies them The number of accidents also increased gradually. The statistics of the World Health Organization point out that about 1.35 billion people die and approximately 50 million road traffic collisions occur every year. One of the common causes that leads to an increase in accidents is driver behavior. It is difficult to deploy in old cars. In addition, wearable devices are disadvantaged by safe driving operations, and the obtained signals can be affected by a few natural structures of the human body. A spectacle with an eye blink sensor is used to detect the driver's drowsiness Objective: The main objective of this project is to provide a study on driver behavior analysis methods. We focus on driver-oriented applications, with three main sub-applications: Accident prevention, Driving style assessment, and Driver intent prediction. The methods we are reviewing in this paper are classified according to their objective which is one of these sub-applications, and their input factors are taken into account during the analysis phase. They can be either quantitative or qualitative factors. The results show that Descriptive statistics and Bayesian classifiers are the methods that were adopted in all three sub-applications and operated on both quantitative and qualitative factors. As for the most employed methods we find the Hidden Markov Model, Support Vector Machine, and Image processing. to develop an intelligent system that will use computer vision approaches for the classification of driver behavior using the Deep Learning Approach. By the use of computer technology, we want to make it simpler to detect the behavior of the driver just from the images. Area: Vehicle-Oriented applications: According to (Meiring, GA M et al., 2015) and (Mittal et al., 2016), the major application areas that fall into this category are "Intelligent Vehicles Systems and Autonomous Vehicles", "Driver Assistance" and "Accidents detection". Intelligent vehicle systems and autonomous vehicles are recent fields of research that aim to automate the functions of cars by exploiting new technologies in communications and data analysis. Google has developed its first fully autonomous car prototype, followed by automotive manufacturers Tesla, Mercedes, and Volkswagen.

This type of research uses advanced vehicular control and environmental detection technologies based on real-time data flow (traffic data, nearby vehicles, etc.). In addition, the connected objects integrated into the vehicle allow continuous communication with various components of the vehicular system. Regarding the automatic detection of accidents, several studies have been carried out in this direction. The major function of this field is the immediate dispatch of emergency and assistance services to the injured driver who may be unconscious and unable to report the accident himself. The techniques used include real-time evaluation of the vehicle's properties (speed, acceleration, sudden stop, etc.), and they enable the identification of abnormal events that are likely to indicate that the vehicle in question has just had an accident. As for driving assistance, it is an application that aims to facilitate the driving task for the driver (parking assistance, video blind spot, etc.).There are nowadays more developed assistance systems introduced by car manufacturers that try to minimize driver error rates due to inattention, distraction, and carelessness. These applications enable effective planning of road maintenance and traffic management. This category also has commercial applications by transport companies; the main objective of fleet management is to control the maintenance of vehicles, monitor their speeds as well as fuel consumption, health, and safety inspection. With effective fleet management, companies can minimize the risks to which vehicles and drivers are subjected, improve the efficiency of their services and reduce overhead costs. Driver-Oriented Applications: This category constitutes the general context of our study. It includes all the research that considers the driver as the main focus, they are represented in the color blue in Figure 1. Driver attention evaluation is one of the main areas of behavioral research, the level of attention is often analyzed by acquisition platforms of the driver's physiological data. These sensors provide information such as eye activity, driver's face tilt, heart rate, and much other information to monitor the somnolence of drivers and their degree of consciousness. As for distraction detection, secondary task recognition systems are developed to identify the degree of driver concentration on the road. They can identify distractions from the driver's reactions. Another research area that we classify in this category is the driving style assessment and driver intent prediction. The first application consists of classifying the driving mode according to several criteria applied to the driver's actions (acceleration, speed, braking, steering, etc.).

The most common styles in scientific literature are the Aggressive style and the Risky style.

These techniques are very useful for automobile insurers who adopt Usage Based Insurance, this technique calculates the insurance costs of each customer according to their driving score and performance. Regarding driver intent prediction, this application consists in predicting the future actions of the driver using the techniques of automatic recognition of maneuvers ALGORITHMS Gabor filter *(for Image Preprocessing):* In image processing, a Gabor filter, named after Dennis Gabor, is a linear filter usedfor texture analysis, which essentially means that it analyzes whether there is any specific frequency content in the image in specific directions in a localized region around the point or region of analysis. Frequency and orientation representations of Gabor filters are claimed by many contemporary vision scientists to be similar to those of the human visual system. *They have been found to be particularly appropriate for texture representation and discrimination.* In the spatial domain, a 2-D Gabor filter is a Gaussian kernel function modulated bya sinusoidal plane wave. A convolutional neural network (CNN) is a special architecture of artificial neural networks, proposed by Yann LeCun in 1988. CNN uses some features of the visual cortex. One of the most popular uses of this architecture is image classification. For example, Facebook uses CNN for automatic tagging algorithms, Amazon — for generatingproduct recommendations, and Google for searching through among users' photos. Let us consider the use of CNN for image classification in more detail. The main task of image classification is acceptance of the input image and the following definition of its class. This is a skill that people learn from their birth and can easily determine that the image in the picture is

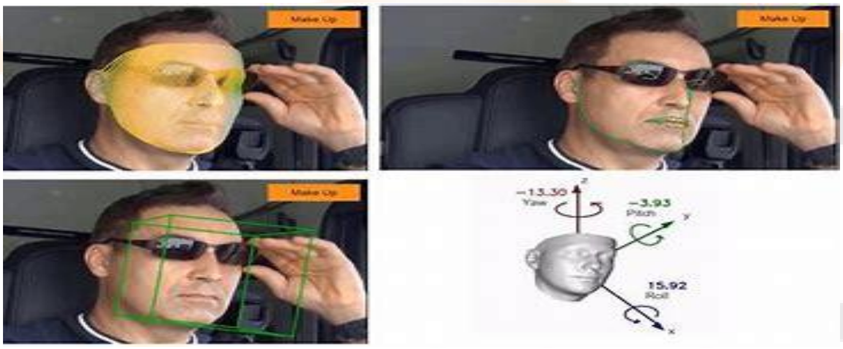

Fig: CNN Image Classification

The Convolution layer is always the first. The image (matrix with pixel values) is entered into it. Imagine that the reading of the input matrix begins at the top left of the image.Next, the software selects a smaller matrix there, which is called a filter (or neuron, or core). Then the filter produces convolution, i.e. moves along the input image. The filter's task is to multiply its values by the original pixel values. All these multiplications are summed up. One number is obtained in the end. Since the filter has read the image only inthe upper left corner, it moves further and further right by 1 unit performing a similar operation. After passing the filter across all positions, a matrix is obtained, but smaller than an input matrix. This operation, from a human perspective, is analogous to identifying boundaries and simple colors on the image. But in order to recognize the properties of a higher level such as the trunk or large ears the whole network is needed. The convolutional layer is the core building block of a CNN, and it is where the majority of computation occurs. It requires a few components, which are input data, a filter, and a feature map. Let's assume that the input will be a color image, which is made up of a matrix of pixels in 3D. This means that the input will have three dimensions—height, width, and depth—which correspond to RGB in an image. We also have a feature detector, also known as a kernel or a filter, which will move across the receptive fields of the image, checking if the feature is present. This process is known as a convolution. The feature detector is a two-dimensional (2-D) array of weights, which represents part of the image. While they can vary in size, the filter size is typically a 3x3 matrix; this also determines the size of the receptive field. The filter is then applied to an area of the image, and a dot product is calculated between the input pixels and the filter. This dot product is then fed into an output array. Afterward, the filter shifts by a stride, repeating the process until the kernel has swept across the entire image. The final output from the series of dot products from the input and the filter is known as a feature map, activation map, or convolved feature. After each convolution operation, a CNN applies a Rectified Linear Unit (ReLU) transformation to the feature map, introducing nonlinearity to the model.

(a) Normal Image (b) After CNN applied

The network will consist of several convolutional networks mixed with nonlinear and pooling layers. When the image passes through one convolution layer, the output of the first layer becomes the input for the second layer. And this happens with every further convolutional layer. The nonlinear layer is added after each convolution operation. It has an activationfunction, which brings nonlinear property. Without this property, a network would not besufficiently intense and will not be able to model the response variable (as a class label).

The pooling layer follows the nonlinear layer. It works with the width and height of theimage and performs a downsampling operation on them. As a result, the image volume isreduced.

This means that if some features (for example boundaries) have already beenidentified in the previous convolution operation then a detailed image is no longerneeded for further processing, and it is compressed to less detailed pictures.

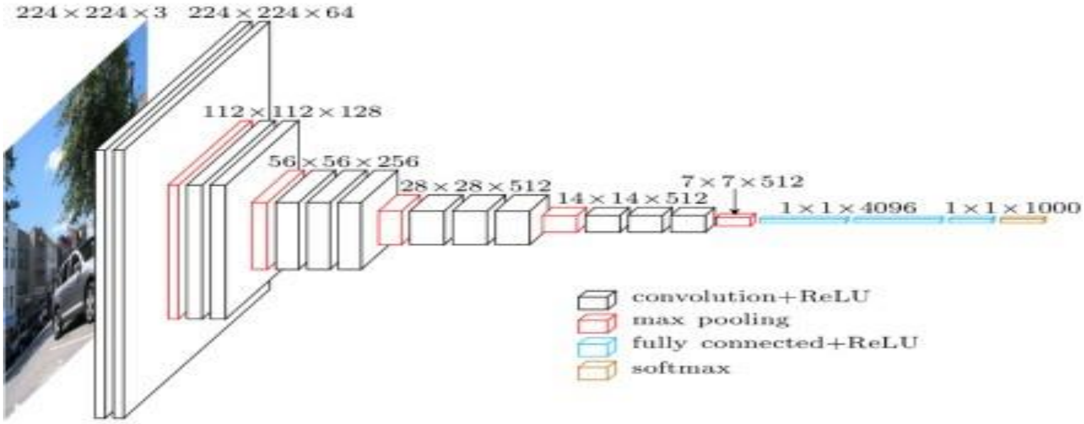

Fig. 4.3: VGG16 Architecture

VGG16 (VGG - Visual Geometry Group): VGG16 is a convolution neural net (CNN) architecture that was used to win ILSVR (Imagenet) competition in 2014. It is considered to be one of the most excellent vision model architectures till date. The most unique thing about VGG16 is that instead of having a large number of hyper-parameter they focused on having convolution layers of a 3x3 filter with a stride 1 and always used the same padding and max pool layer of a 2x2 filter of stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end, it has 2 FC(fully connected layers) followed by a softmax for output. The 16 in VGG16 refers to it has 16 layers that have weights. This network is pretty large and it has about 138 million (approx) parameters. Fig. 4.3: VGG16 ArchitectureOn the convolutional output, we take the first 2 x 2 region and calculate the max value from each value in the 2 x 2 block. This value is stored in the output channel, which makes up the full output from this max pooling operation. We move over by the number of pixels that we defined our stride size to be. We're using 2 here, so we just slide over by 2, then do the same thing. We calculate the max value in the next 2 x 2 block, store it in the output, and then, go on our way sliding over by 2 again. Once we reach the edge over on the far right, we then move down by 2 (because that's our stride size), and then we do the same exact thing of calculating the max value for the 2 x 2 blocks in this row. After performing max pooling, we can see the dimension of this image was reduced by a factor of 2 and is now 13 x 13. The primary task of a Deep Neural Network – especially in the case of Image recognition, Video Processing, etc is to extract the features systematically by identifying edges and gradients and forming textures on top of it. As whole, convolutional layers in the Deep Neural Networks form parts of objects and finally objects which can summarize the features in an input image. In this process, maintaining the same image size throughout the Neural Network will lead to the stacking of multiple layers. This is not sustainable due to the huge computing resources it demands. At the same time, we need enough convolutions to extract meaningful features.

Proposed Architecture:

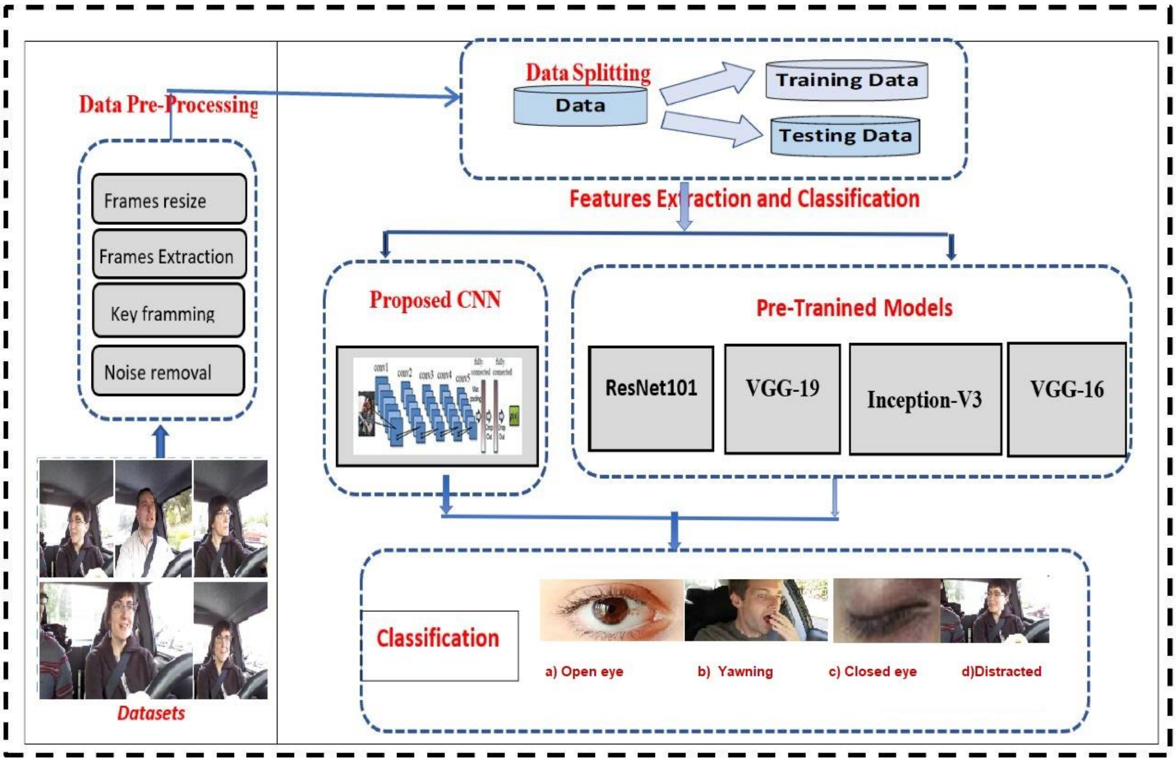

Fig: Proposed Architecture

A proposed system for driver behavior detection using a Convolutional Neural Network (CNN) model involves the use of advanced machine learning algorithms to detect and analyze driver behavior patterns. The CNN model is a deep learning algorithm that is specifically designed for image recognition and analysis. This makes it an ideal tool for analyzing video footage from cameras mounted inside vehicles to detect and classify driver behaviors.

The proposed system would consist of several components, including a camera system, a data storage and processing unit, and a CNN model. The camera system would be mounted inside the vehicle and would capture video footage of the driver and their surroundings. The data storage and processing unit would store and process this video footage, preparing it for analysis by the CNN model.

The CNN model would be trained on a large dataset of labeled driver behavior patterns, such as aggressive driving, distracted driving, drowsy driving, and so on. Once trained,

the model would be able to detect these behaviors in real time by analyzing the video footage captured by the camera system. The output of the CNN model could then be used to trigger alerts or warnings to the driver or other systems in the vehicle.

VGG16 is a convolutional neural network architecture in that the network consists of 16 layers comprising mainly of convolutional layers followed by some fully connected layers. It has a simple yet effective design, where each layer of the network comprises 3x3 convolutional filters with a stride of 1 and a padding of 1. Max-pooling layers are used after every two convolutional layers, thereby reducing the spatial dimensionality of the output of the layers.

On the other hand, ResNet101 is a deep residual network architecture. It consists of 101 layers and has a unique design that helps to address the problem of vanishing gradients. The network uses a residual block that enables the network to learn an identity function, which helps in maintaining the gradients throughout the network. The residual block contains a shortcut connection that skips one or more layers, allowing for easier training of deeper models.

Understanding the architecture and design of VGG16:

Understanding VGG16's architectural patterns, the role of convolutional filters, and how Max-pooling layers are used after specific layers to reduce the spatial dimensions of the output.

Understanding the architecture and design of ResNet101:

Understanding ResNet101's unique design that consists of residual blocks that helps in addressing the problem of vanishing gradients, and how the identity function helps in maintaining the gradients throughout the network.

The proposed system has several potential benefits, including improved safety on the roads, reduced accidents caused by driver error, and increased efficiency in transportation systems. It could also be used for research purposes to better understand driver behavior patterns and develop more effective interventions to improve road safety.

VGG-16, VGG-19, and ResNet101 are two popular models in the field of Machine Learning for computer vision tasks such as image classifications, object detections, etc.

WORKFLOW

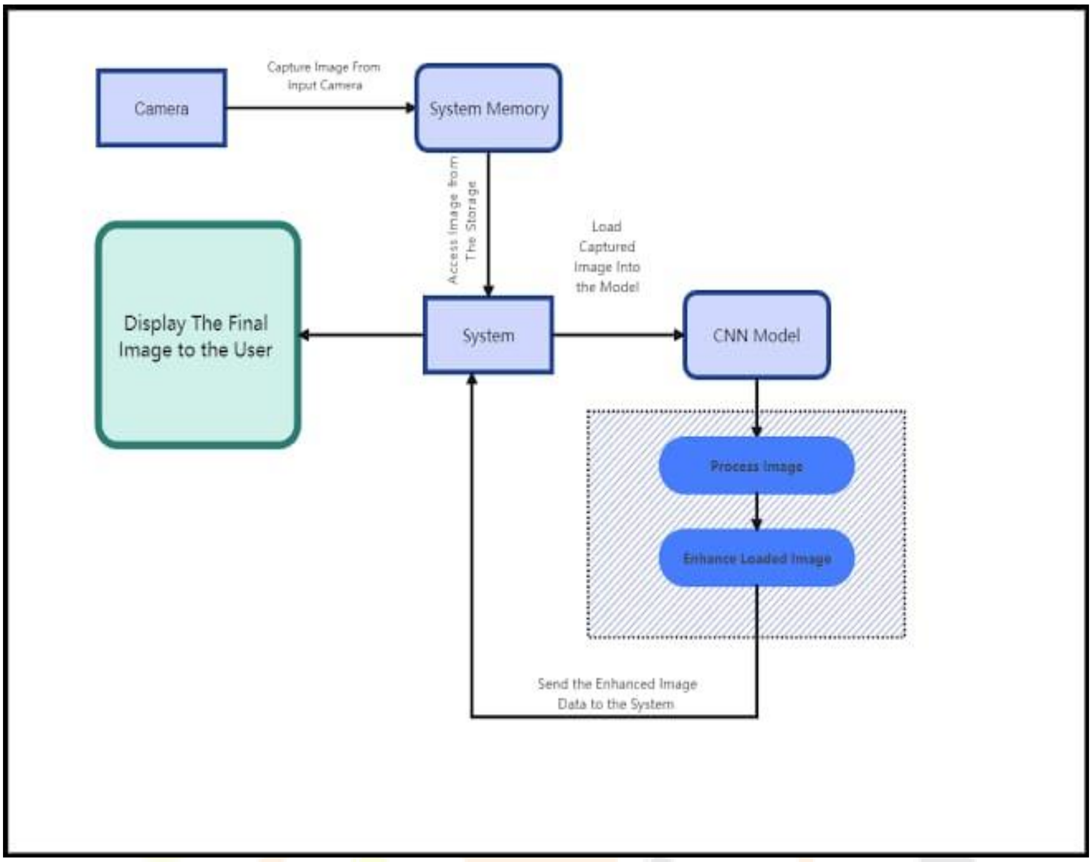
Camera
Capture Image From Input Camera
System Memory
Access Image from The Storage
Display The Final Image to the User
System
Load Captured Image Into the Model
CNN Model
Process Image
Enhance Loaded Image
Send the Enhanced Image Data to the System

Fig 4.4: Workflow model of the proposed architecture

Workflows are a series of steps that need to be completed in a process. Think of it literally as work flowing from one stage to the next, whether that's through a colleague, tool, or another process.

The workflow modal of chauffeurs' behavior recognition based on CNN (Convolutional Neural Network) is an innovative system that aims to identify and analyze chauffeurs' behaviors using advanced technology. The workflow modal of behavior recognition based on CNN is a complex system that involves multiple stages, each with its unique functionalities. The system works by capturing driving data using sensors and cameras installed in the vehicle. This data is then pre-processed and analyzed using machine learning techniques to identify specific behavior patterns.

Data Acquisition

The first step in the workflow modal of the chauffeur's behavior recognition is data acquisition. This involves gathering data from various sources, such as cameras and sensors, installed in the vehicle. The data collected includes various parameters, such as speed, acceleration, and brake force, which are used to train the neural network.

Data Pre-processing

The next step is data pre-processing, where the raw data collected is cleaned and pre-processed to remove any inaccuracies or errors. This step involves filtering data, removing outliers, and smoothing the data to improve the accuracy of the neural network.

Feature Extraction

In this step, the preprocessed data is analyzed to extract relevant features that are used to train the neural network. Feature extraction involves identifying specific patterns in the data that are indicative of particular behaviors, such as aggressive driving or distraction.

Neural Network Training

Once the relevant features have been extracted, the data is used to train the neural network. This involves using a convolutional neural network (CNN) to classify the behavior patterns identified in the data. The neural network is trained using both labeled and unlabelled data, with the labeled data used to teach the network what specific behaviors look like.

Behavioral Analysis

The final step involves using the trained neural network to analyze real-time driving data and identify specific behavior patterns in the data. This analysis can be used to provide real-time feedback to the driver, such as warnings or alerts, to prevent accidents or improve driving habits.

Conclusion:

CNN models have shown high accuracy in recognizing various driving behaviors such as lane departure, hard braking, and distracted driving, among others. By using deep learning algorithms, CNN models can analyze large amounts of data from various sensors such as accelerometers, gyroscopes, and cameras, and accurately distinguish between safe and risky driving behaviors. This can help prevent accidents and reduce the number of fatalities on the road. Training the CNN model requires a large amount of labeled data that can be used to classify different driving behaviors. The data can be gathered from various sensors such as cameras and accelerometers and then fed into the CNN model to train it. The model is then fine-tuned through iterations until it accurately classifies various driving behaviors.

Testing and Validation, after training the CNN model, it needs to be tested and validated on new data to see how well it performs in classifying new driving behaviors. This ensures that the model is reliable and accurate and can be used in real-world applications.

Deployment Once the model has been trained and validated, it can be deployed in real-world scenarios such as in cars or trucks. It can be used to alert drivers when they engage in risky behavior, or it can notify fleet managers of risky driving in commercial vehicles.

Grey scaling is used to convert the images into 0's AND 1's In grayscale, each pixel of the image is represented by a single value that represents the intensity of the light in that pixel, and its represented by a single value between 0 and 255, where 0 represents black, grayscale is a useful technique in machine learning for converting color images to simpler and easier-to-process grayscale images. It reduces dimensionality while retaining important features of the image, making it easier for machine learning algorithms and analyze data. In our model we have also deployed yawning and drowsiness detection, a buzzer will be beeped when the driver is tending to sleep, we are able to achieve 91% of accuracy in these model, In this model, we are able to detect the driver's behavior they are observed and the output will be printed on the screen saying the particular action being done by the driver. the only drawback in this model is that we have used pictures as our data sets and fed the pictures which are inside a car so able to detect only when the person is inside the car and proper lighting should be present to detect the chauffeur’s behavior.

References

1] (2019). The National Highway Traffic Safety Administration (NHTSA). Accessed: Aug. 5, 2020. [Online]. Available: https://www.nhtsa. gov/technology-innovation/automated-vehicles-safety

[2] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Expression- invariant 3D face recognition," in Audio-Video-Based Biometric Person Authentication (Lecture Notes in Computer Science), vol. 2688, J. Kittler and M. Nixon, Eds. Guildford, U.K. Berlin, Germany: Springer, Jun. 2003, pp. 62–70.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in Proc. Brit. Mach. Vis. Conf., Swansea, U.K., Sep. 2015, pp. 1–12.

[4] L. De Silva, T. Miyasato, and R. Nakatsu, “Facial emotion recognition using multi-modal information,” in Proc. Int. Conf. Inf., Commun. Signal Process. (ICICS), Singapore, vol. 1, 1997, pp. 397–401.

[5] S. Jha and C. Busso, "Estimation of gaze region using two-dimensional probabilistic maps constructed using convolutional neural networks," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Brighton, U.K., May 2019, pp. 3792–3796.

[6] N. Li and C. Busso, “Calibration free, user-independent gaze estimation with tensor analysis,” Image Vis. Comput., vol. 74, pp. 10–20, Jun. 2018.

[7] S. Jha and C. Busso, “Challenges in head pose estimation of drivers in naturalistic recordings using existing tools,” in Proc. IEEE Int. Conf.

Intell. Transp. (ITSC), Yokohama, Japan, Oct. 2017, pp. 1624–1629.

[8] S. Jha and C. Busso, “Probabilistic estimation of the driver’s gaze from head orientation and position,” in Proc. IEEE 20th Int. Conf. Intell.

Transp. Syst. (ITSC), Yokohama, Japan, Oct. 2017, pp. 1630–1635.

[9] S. Jha and C. Busso, “Probabilistic estimation of the gaze region of the driver using dense classification,” in Proc. 21st Int. Conf. Intell. Transp.

Syst. (ITSC), Maui, HI, USA, Nov. 2018, pp. 697–702.

[10] S. Vora, A. Rangesh, and M. M. Trivedi, “Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis,” IEEE Trans. Intell. Veh., vol. 3, no. 3, pp. 254–265, Sep. 2018.

[11] S. Jha and C. Busso, “FI-CAP: Robust framework to benchmark head pose estimation in challenging environments,” in Proc. IEEE Int. Conf.

Multimedia Expo (ICME), San Diego, CA, USA, Jul. 2018, pp. 1–6.

[12] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye, and iris for robust gaze zone estimation," in Proc. IEEE Conf. Intell. Transp. Syst., Qingdao, China, Oct. 2014,

pp. 988–994.

[13] B. Vasli, S. Martin, and M. M. Trivedi, “On driver gaze estimation:

Explorations and fusion of geometric and data-driven approaches," in Proc. 19th IEEE Int. Conf. Intell. Transp. Syst., Rio de Janeiro, Brazil, Nov. 2016, pp. 655–660.

[14] S. Jha and C. Busso, “Analyzing the relationship between head pose and gaze to model driver visual attention,” in Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC), Rio de Janeiro, Brazil, Nov. 2016, pp. 2157–2162.

[15] G. Li et al., “Drivers’ visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China,” J. Saf. Res., vol. 71, pp. 219–229, Dec. 2019.

[16] G. Li et al., “Influence of traffic congestion on driver behavior in post-congestion driving," Accident Anal. Prevention, vol. 141, Jun. 2020, Art. no. 105508.

[17] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “ELAN: A professional framework for multimodality research,” in Proc. Int. Conf. Lang. Resour. Eval. (LREC), Genoa, Italy, May 2006, pp. 1556–1559.

[18] F. Yu et al., “BDD100K: A diverse driving dataset for heterogeneous multitask learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 2633–2642.

[19] D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, “Predicting driver attention in critical situations,” in Proc. Asian Conf. Comput. Vis. (ACCV), in Lecture Notes in Computer Science, vol. 11365, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., Perth, WA, Australia. Berlin, Germany: Springer, Dec. 2018, pp. 658–674.

[20] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in Proc. ICCV, Venice, Italy, Oct. 2017, pp. 5000–5009.

[21] P. Sun et al., “Scalability in perception for autonomous driving: Waymo open dataset,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 2443–2451.

[22] H. Caesar et al., “Nuscenes: A multimodal dataset for autonomous driving,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Aug. 2020, pp. 11618–11628.

[23] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, “Predicting the driver’s focus of attention: The DR(eye)VE project,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1720–1733, Jul. 2019.