EJECE, European Journal of Electrical Engineering and Computer Science
ISSN: 2736-5751

Convolutional Neural Networks (CNN) based Eye-Gaze Tracking System using Machine Learning Algorithm

Prakash Kanade, Fortune David, and Sunay Kanade

***Abstract* — To avoid the rising number of car crash deaths, which are mostly caused by drivers' inattentiveness, a paradigm shift is expected. The knowledge of a driver's look area may provide useful details about his or her point of attention. Cars with accurate and low-cost gaze classification systems can increase driver safety. When drivers shift their eyes without turning their heads to look at objects, the margin of error in gaze detection increases. For new consumer electronic applications such as driver tracking systems and novel user interfaces, accurate and effective eye gaze prediction is critical. Such systems must be able to run efficiently in difficult, unconstrained conditions while using reduced power and expense. A deep learning-based gaze estimation technique has been considered to solve this issue, with an emphasis on WSN based Convolutional Neural Networks (CNN) based system. The proposed study proposes the following architecture, which is focused on data science: The first is a novel neural network model that is programmed to manipulate any possible visual feature, such as the states of both eyes and head location, as well as many augmentations; the second is a data fusion approach that incorporates several gaze datasets. However, due to different factors such as environment light shifts, reflections on glasses surface, and motion and optical blurring of the captured eye signal, the accuracy of detecting and classifying the pupil centre and corneal reflection centre depends on a car environment. This work also includes pre-trained models, network structures, and datasets for designing and developing CNN-based deep learning models for Eye-Gaze Tracking and Classification.**

***Index Terms* — Eye Gaze tracking, Convolutional Neural Network (CNN), LeenaBOT.**

I. Introduction

Driver inattention and glance diversion from the lane are the primary causes of traffic collisions [1]. Driver diversion is the most common cause of focus deviation from the lane, and it can put drivers, riders, and pedestrians in grave danger. According to the US Department of Transportation, distracted drivers killed 3300 people and wounded 5,29,000 people in 2018 [2]. Distracted driving may occur because of any behaviour that diverts the driver's attention away from the primary task of driving.

It can occur for a variety of purposes, but the most important include using a cellphone, monitoring the radio, eating and drinking, and using a global positioning system (GPS). According to the National Highway Traffic Safety Administration (NHTSA), using a cell phone while driving raises the risk of a car accident by three times [3]. Using a mobile phone allows drivers to take their eyes off the road for the longest amount of time (EOR). In short, it can be a source of driver distraction, and driver look tracking technologies can play a critical role in preventing car accidents. In the pursuit of crash avoidance, the classification of driver gaze focus is becoming increasingly important.

Eye tracking research has gradually been applied in a variety of applications, including driving fatigue alert systems, mental health screening, an eye-tracking powered wheelchair, and other human–computer interface systems. However, there are many limitations, including dependable real-time performance, high precision, device availability, and a lightweight and non-intrusive device. It is also critical to improve device robustness in the face of obstacles including shifting lighting conditions, physical eye shape, surrounding eye characteristics, and eyeglass reflections.

Several similar studies have suggested eye-controlled wheelchair systems; however, these seldom discuss the device's machine efficiency limitations, physical and surrounding problems outside the system, algorithm novelty, and overall user comfort and protection. Convolutional Neural Networks (CNNs) are also a cutting-edge and effective method for solving computationally and data-intensive problems. CNN is a leader in a wide range of technologies, including object classification, speech recognition, natural language processing, and even wheelchair control; a more in-depth overview of the literature will be covered in the following pages. However, the paper lacks high precision, real-time implementation, and the specifics of such a design, which may be useful for future improvement [4]. Even though many experiments on gaze engagement have been published, the efficiency of these approaches in smart interactive settings under real-world conditions is still lacking.

A deep learning-based gaze estimation technique has been considered to solve this challenge, with an emphasis on Convolutional Neural Networks (CNN) based methods. The proposed study proposes the following architecture, which is focused on data science: The first is a novel neural network model that is programmed to manipulate any possible visual feature, such as the states of both eyes and head location, as well as many augmentations; the second is a data fusion approach that incorporates several gaze datasets. This project also includes pre-trained models, network structures, and datasets for designing and developing CNN-based deep learning models for eye-gaze tracking and classification.

Submitted on March 12, 2021.
Published on April 07, 2021.

Prakash Kanade, Hobby Researcher in Robotics, Artificial Intelligence, IOT, USA.
(e-mail: Prakashsrs@gmail.com)

Fortune David, Faculty, Facilitator, LeenaBOT Robotics LLC, USA.
(e-Email: fortunekbz2009@gmail.com)
Sunay Kanade, Student, LeenaBOT Robotics LLC, USA.
(e-mail: sunaykanade6@gmail.com)

DOI: http://dx.doi.org/10.24018/ejece.2021.5.2.314

II. Literature Survey

Choi et al. [5] for driver gaze zone description and head-pose estimate, a five-layered CNN-based methodology was suggested. They created a dataset that included photographs of male and female drivers, as well as drivers who wore eyeglasses. They developed a CNN model based on the dataset that can identify 9 gaze zones of drivers and approximate their head-pose. The left mirror, right mirror, rearview mirror, steering, gear, center, left windscreen, and right windscreen are all represented by separate gaze zones in a vehicle.

Naqvi et al. [6] a CNN-based model using a near-infrared (NIR) camera that considers head and eye motions while not obstructing drivers' vision took on this challenge. One NIR sensor, one zoom lens, and six NIR light-emitting diodes (LEDs) are used to illuminate the device. The frontal view image of the driver is captured by the NIR camera and then sent to a laptop via a USB interface cable.

Konrad et al. [7] introduced a CNN-based end-to-end gaze estimating methodology for near-eye displays to solve this issue. They created a dataset of eye photographs of people staring at different calibration points on a tablet. A camera was mounted very close to the faces of the subjects in order to catch the photographs. They created a basic CNN model (using the LeNet architecture) based on the dataset, which takes user photos as input and estimates the users' gaze orientation based on the x and y coordinates on the screen. The authors approached the gaze estimation problem as a multi-class classification problem, with each class represented by a screen point. On the captured dataset, the technique created an angular error of 6.7 degrees, which is unacceptable. Bad image quality (2828) and dataset heterogeneity used to train the network are likely to blame for the network's poor results. Just 5 topics are represented in the dataset.

Hickson et al. [8] eye trackers were used to suggest a method for classifying facial expressions in VR systems. The methodology analyses only a partly occluded face of participants as they are immersed in a virtual reality experience to automatically infer facial expressions. Images of the user's eye was taken by a gaze monitoring camera within a headset and used to infer a subset of facial expressions in the analysis. The photos are used to create real-time animated avatars that function as a proxy for users' expressive abilities. They presented a new method for optimizing deep CNN model accuracy. The methodology was tested and found to have a mean accuracy of 74%.

Garbin et al. [9] introduced a large-scale dataset for VR device model training. The eye photographs were captured using a virtual reality head-mounted device of two paired eyes facing cameras. The photographs in the dataset are split into four subsets and were taken from the eye regions of 152 subjects. The first subset contains 12,759 images that have been labelled, while the second subset contains 252,690 images that have not been labeled. The third subset includes 91,200 frames chosen at random from 1.5-second video sequences, and the final subset includes 143 pairs of left and right point cloud data. An experiment was conducted to assess the dataset's accuracy, and the findings suggest that it can be used to create eye-tracking models for VR applications.

Alsharif et al. [10] using the LSTM network, another gaze-based typing approach was added. Input data was used to train an LSTM network in this process. The Connectionist Temporal Classification (CTC) loss function was used during training to enable the network to produce characters without using HMM states. The network will use this feature to map long input sequences to short output sequences. The network generates a matrix that corresponds to the set of permissible characters after preparation. A Finite State Transducer was used to restrict the processing to a certain collection of terms. The system has been tested, and the findings indicate that it has a 92 percent classification accuracy.

Liang et al. [11] introduced an eye-tracking-based video-based recognition model for biometric systems. They made video clips for various subjects to watch in order to collect eye-tracking data that reflected their physiological and behavioural characteristics, such as acceleration, muscle, and geometric features. These characteristics are derived from the eye gaze data and used as biometric features to classify people. They choose related features for biometric authentication using a feature selection algorithm (based on shared knowledge of features). The features that were chosen were then used to train two classifiers, NN and SVM. The findings of the experiments demonstrate that using video-based eye-tracking data to measure biometric data is a viable option.

Shin et al. [12] introduced a tool for estimating gaze depth in VR and AR systems. They collected data from two eyes using a binocular eye tracker, which included pupil center, gaze orientation, and inter pupil time. These characteristics are then used to create a neural network (NN) model for predicting gaze depth. A dataset of photographs from 13 subjects was used to test the process. Specific models were created for each of the 13 subjects in the evaluation, as well as a standardized model for the 13 subjects. The system provided a gaze depth accuracy of 90.1 percent for individual models and 89.7 percent for the generalized model, according to the experimental results.

Wang et al. [13] introduced a gaze estimation methodology based on CNN and random forest regression to solve this problem. They developed a hybrid methodology for studying CNN-based image features (called deep features) for gaze estimation instead of hand-crafted features. They specifically trained a CNN model on various eye images and extracted the contribution of the network's last fully linked layer. Following that, the CNN-based features were used to train a random forest regressor to learn mappings between deep features and gaze coordinates. With a prediction error of 1.530, the technique was effective. When opposed to hand-engineered features, the findings reveal that deep features significantly increase performance on regression-based algorithms.

Palmero et al. [14] in a remote camera, we tackled the issue of head-pose and person-independent 3D gaze estimation. They developed a system for modelling both appearance- and shape-based gaze prediction cues. The full-face and eye photographs reflect appearance-based cues, while 3D facial landmarks obtained from a 68-landmark model that models the global outline of the face represent shape-based cues. An RCNN is jointly trained using appearance-based and shape-based features.

III. Problem Statement

- Driver diversion is the most common cause of focus deviation from the lane, and it can put drivers, riders, and pedestrians in grave danger. Eye monitoring methods have been gradually applied in driving fatigue-warning systems as a result of research.
- However, there are many limitations, including dependable real-time performance, high precision, device availability, and a lightweight and non-intrusive device.
- Even though many experiments on gaze engagement have been published, the efficiency of these approaches in smart interactive settings under real-world conditions is still lacking.
- A deep learning-based gaze estimation technique has been considered to solve this challenge, with an emphasis on Convolutional Neural Networks (CNN) based methods.

IV. Objectives

- The use of Convolutional Neural Networks (CNN)-based deep learning-based gaze inference techniques has been considered.
- To integrate pre-trained models, network structures, and datasets that can be used to build and improve CNN-based deep learning models for training.
- GazeCapture uses binary images to train a deep neural network and estimate the user's gaze orientation.

V. Methodology

The technique used for the planned project is broken down into three stages:

A. Data Preprocessing

Preprocessing methods that GazeCapture dataset followed before being used as the model's input is shown in Fig. 1.

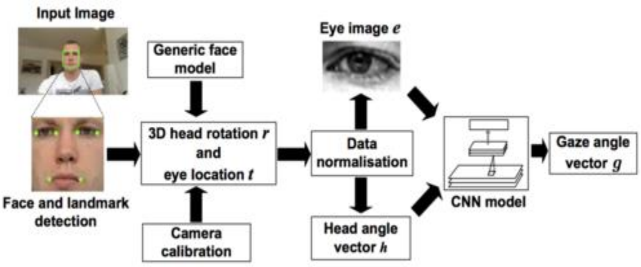

Fig. 1. Overall Proposed Preprocessing.

The function extractor is made up of a hierarchy of CNNs that have been trained separately. One CNN serves as an eye area classifier, while the others estimate attribute positions explicitly. If all corneal reflections are clear and the pupil is not significantly occluded, the classifier network concludes that the pupil is not significantly occluded (such as when a user blinks). Several independent position prediction networks decide the direction of the pupil centre and corneal reflections if this network decides that the eye area is valid. As seen in Figure 1, the classifier and location prediction networks have identical base CNN architectures. Several convolutional layers (with batch normalization and optional pooling) are accompanied by dense layers that are totally linked.

B. Head Pose Estimation

Head posture may be an extra aspect to the training model due to the user's unlimited head movement. As a result, the Euler's angles: pitch, yaw, and roll, as seen in Fig. 2, are used to extract the head pose's orientation.

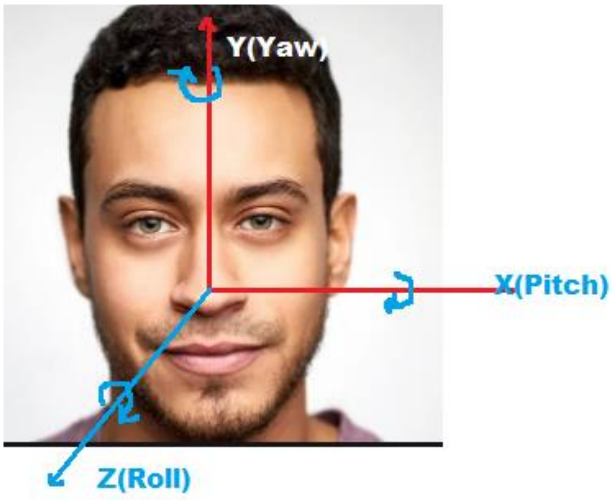

Fig. 2. Head Pose Euler's Angles, Pitch, yaw, and roll.

The presence of a face and the precise coordinates of the eye regions are determined using a commercial head tracker. When the function extractor is first started or when local monitoring fails, these regions are used to initialize it. As a result, the machine only employs the head sensor on rare occasions. The regions that initialize the feature extractor are based on the calculation of the previous center of the pupil when the eye features were successfully monitored in the previous frame. A hierarchy of tiny effective CNNs is used to localize narrow eye area windows.

C. Residual Neural Network

The architecture of the training model using Residual Neural Networks as seen in Fig. 3. (ResNet).

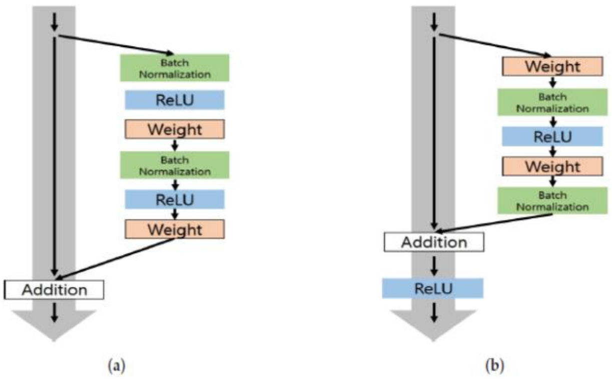

Fig. 3. a) Proposed Architecture; b) Original Architecture.

Training Details:

Metrics: The GazeCapture dataset includes this metric.

The Euclidean interval between the projected and the target is the loss function.

Validation: A small percentage of the data is used to create a validation dataset.

Learning rate: Getting the learning process started.

Adams's optimizer is the name of the optimizer.

Epochs are the number of years in two epochs.

Our training dataset consists of front-facing camera infrared photographs of eye regions with labelled locations of eye features. We had to create a new dataset because one of this kind did not exist publicly. The relative location and orientation differences between the device and the subject's face cause the picture uncertainty we expect in a camera context. Because of the (possibly) lower contrast between the pupil and the iris, which is exacerbated by limited lighting capacity and a noisier camera sensor, there may be more uncertainty.

We gathered and labelled 50 face photographs (40 eye regions) from ten participants who were not wearing glasses (contacts were not controlled for) while using our prototype infrared smartphone. Each participant was instructed to place the computer in a comfortable location and look at a random target on the screen. A single picture was taken by the unit. Before the next image was taken, the subject was allowed five seconds to reposition the computer and replicate the process. The instrument's placement and repositioning culminated in a number of relative distances and orientations between the device and the heads of the participants. For each subject, this procedure was repeated ten times, yielding ten infrared photographs of their face captured by the front-facing camera. A random set of 5 frame indexes is created during the 10-image selection process to artificially increase the amount of invalid eye regions in the dataset. When recording these individual frames, one or both infrared LEDs are disabled (at random), resulting in eye regions that are unsuitable for the gaze-estimation model used in this study. For the remaining 09 frames, all LEDs were turned on. The sample selection app did the orchestration of the LEDs in this manner automatically. Data collected during the eye-experimental tracker's testing was used to validate the networks.

VI. Conclusions

We propose a CNN-based approach for driver gaze classification in the vehicular environment. Face, left eye, and right eye images will be extracted from the input image based on the ROI (region-of-interest) identified by facial landmarks from the facial attribute tracker for driver gaze classification. Further fine tuning with a pre-trained CNN model using the VGG-face network to obtain the appropriate gaze features from the completely linked layer of the network separately for the extracted cropped images of face, left eye, and right eye.

To achieve the final classification result, three distances depending on all of the obtained features are combined. The influence of the CNN paradigm on gaze classification will also be investigated. The studies showed that in the presence of relative motion between the subject and the device, high-quality gaze estimates can be generated when robust eye features are given to the 3D gaze-estimation model. The experiments in this paper were designed to demonstrate how well low-cost CNNs can deliver reliable eye features in the presence of greater relative motion.

Machine-learning algorithms are used in the eye-tracking method to increase the robustness and precision of eye-feature estimation (the location of corneal reflections and the center of the pupil). The eye characteristics are then used by a gaze-estimation algorithm that is unaffected by the subject's and smartphone's relative motion. Our hybrid eye-tracking technology (which benefits from infrared illumination) achieves considerably higher accuracy (more than 100 percent) than previous systems that estimate gaze location using natural light and appearance-based approaches.

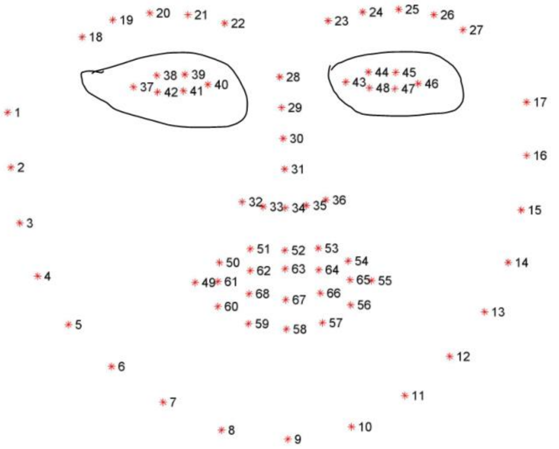

Fig. 4. Facial Landmarks to detect eye.

The hybrid approach, which includes a machine-learning feature extraction stage followed by a geometric 3D model, improves results. This advancement could make mobile apps that involve visual scanning behaviour analysis of subjects viewing a small number of objects on a smartphone screen possible.

VII. Future Enhancement

Experiments with the free Columbia gaze dataset CAVE-DB showed that our approach can be extended to photographs taken with a visible light camera without the use of an extra illuminator.

VIII. Application

Eye tracking research has gradually been applied in a variety of applications, including driving fatigue alert systems, mental health screening, an eye-tracking powered wheelchair, and other human–computer interface systems like AI added medical treatment systems and remote treatment systems [15], [16].

References

[1] Andronicus A. Akinyelu and Pieter Blignaut, "Convolutional Neural Network-Based Methods for Eye Gaze Estimation: A Survey," IEEE Access, pp. 142581-142605, Volume 8, 2020.

[2] Braiden Brousseau, Jonathan Rose and Moshe Eizenman, "Hybrid Eye-Tracking on a Smartphone with CNN Feature Extraction and an Infrared 3D Model," Sensors 2020, Volume 20, Issue 543, pp. 1-21, 2020.

[3] En Teng Wong, Seanglidet Yean, Qingyao Hu, Bu Sung Lee, Jigang Liu, Rajan Deepu, “Gaze Estimation Using Residual Neural Network,” IEEE Xplore, PerCom Work in Progress on Pervasive Computing and Communications, 2019.

[4] Joseph Lemley, Anuradha Kar, Alexandru Drimbarean and Peter Corcoran, “Convolutional Neural Network Implementation for Eye-Gaze Estimation on Low-Quality Consumer Imaging Systems,” IEEE Transactions on Consumer Electronics, 2018.

[5] H. Choi, Y. G. Kim, and T. B. H. Tran, ‘‘Real-time categorization of driver’s gaze zone and head pose -using the convolutional neural network,’’ in Proc. HCI Korea, Jan. 2016, pp. 417–422.

[6] R. Naqvi, M. Arsalan, G. Batchuluun, H. Yoon, and K. Park, ‘‘Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor,’’ Sensors, vol. 18, no. 2, p. 456, Feb. 2018.

[7] R. Konrad, S. Shrestha, and P. Varma, ‘‘Near-eye display gaze tracking via convolutional neural networks,’’ Standford Univ., Standford, CA, USA, Tech. Rep., 2016.

[8] S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. Essa, ‘‘Eyemotion: Classifying facial expressions in VR using eye-tracking cameras,’’ in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2019, pp. 1626–1635.

[9] S. J. Garbin, O. Komogortsev, R. Cavin, G. Hughes, Y. Shen, I. Schuetz, and S. S. Talathi, ‘‘Dataset for eye tracking on a virtual reality platform,’’ in Proc. Symp. Eye Tracking Res. Appl., Jun. 2020, pp. 1–10.

[10] O. Alsharif, T. Ouyang, F. Beaufays, S. Zhai, T. Breuel, and J. Schalkwyk, ‘‘Long short term memory neural network for keyboard gesture decoding,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2015, pp. 2076–2080.

[11] Z. Liang, F. Tan, and Z. Chi, ‘‘Video-based biometric identification using eye tracking technique,’’ in Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC), Aug. 2012, pp. 728–733.

[12] C. Shin, G. Lee, Y. Kim, J. Hong, S.-H. Hong, H. Kang, and Y. Lee ‘‘Evaluation of gaze depth estimation using a wearable binocular eye tracker and machine learning,’’ J. Korea Comput. Graph. Soc., vol. 24, no. 1, pp. 19–26, 2018.

[13] Y. Wang, T. Shen, G. Yuan, J. Bian, and X. Fu, ‘‘Appearance-based gaze estimation using deep features and random forest regression,’’ Knowl-Based Syst., vol. 110, pp. 293–301, Oct. 2016.

[14] C. Palmero, J. Selva, M. A. Bagheri, and S. Escalera, ‘‘Recurrent CNN for 3D gaze estimation using appearance and shape cues,’’ 2018, arXiv:1805.03064. [Online]. Available: http://arxiv.org/abs/ 1805.03064.

[15] Prakash Kanade, Sunay Kanade, "Medical Assistant Robot ARM for COVID-19 Patients Treatment – A Raspberry Pi Project," International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 10, Pages. 105-111, 2020.

[16] Prakash Kanade, Monis Akhtar, Fortune David, "Computer Networking and Technology Improvement in the Age of COVID-19," International Journal of Advanced Networking and Applications (IJANA), vol. 12, no. 03, Pages. 4592-4595, 2020.

DOI: http://dx.doi.org/10.24018/ejece.2021.5.2.314