ISSN: 2508-7894 © 2022 KODISA & KAIA.
KJAI website: http://acoms.kisti.re.kr/kjai
doi: http://dx.doi.org/10.24225/kjai.2022.9.1.1

A study on driver state recognition using CNN-based multimodal multi-input learning*

Sooah SHIN[1], Jisu KANG[2], Sooah KIM[3], Hwanseo YEO[4], Dongyeon LEE[5], Jeongjyun LEE[6], Gijun HAN[7], Jinho HAN[8]

Received: January 15, 20xx. Revised: November 29, 20xx. Accepted: December 05, 20xx.

Abstract

Accurately identifying the driver's driving status is very important for the safety of the driver and passengers, and many studies are being conducted on this. Many researchers have shown that attempts to identify the driver's status by creating an image of the driver's face and upper body using a camera and then learning CNN-based neural networks are somewhat effective. In addition, attempts have been made to identify the driver's fatigue using electroencephalogram (EEG), heart rate (ECG), and electrooculogram (EOG) data measurements, multiple sensors, and the RNN-LSTM model method. Nowadays, increasing accuracy is essential using simple sensor devices and simple artificial intelligence structures that can be installed in small devices such as on-device AI. In this paper, we used the driver's facial expression and upper body and the sound data generated during driving to recognize the driver's status more accurately. We then showed through experiments that CNN-based neural network learning alone using triple input elements improved meaningful accuracy. To apply on-device AI, we propose a CNN with a simple structure that can collect data using only a camera and a recorder. We compare the proposed method with learning in ResNet50 and Xception to show that it works well. These experimental results show that CNN can be used in multimodal applications and can be an efficient choice over other complex neural network learning methods that use multimodal learning data.

Keywords : Artificial Intelligence, Multimodal, Multi Input, On-device AI, CNN, Driver State Recognition

Major Classification Code : Computers(L63), Computer Software (L86), Technological Change (O33)

1. Introduction

Understanding the driver's emotions or state while driving is very important for the safety of the driver and

* This study was supported by the University Innovation Support Project of Korean Bible University in 2024.

1 First Author, Department of Computer Software, Korean Bible University, Republic of Korea, Email: mojasuah@gmail.com
2 Second Author, Department of Computer Software, Korean Bible University, Republic of Korea, Email: ji_su217_@naver.com
3 Third Author, Department of Computer Software, Korean Bible University, Republic of Korea, Email: izoasooa@naver.com
4 Fourth Author, Department of Computer Software, Korean Bible University, Republic of Korea, Email: hwanseo05@bible.ac.kr
5 Fifth Author, Department of Computer Software, Korean Bible University, Republic of Korea, Email: spy35@bible.ac.kr
6 Sixth Author, Department of Computer Software, Korean Bible University, Republic of Korea, Email: jh082393@naver.com
7 Seventh Author, Department of Computer Software, Korean Bible University, Republic of Korea, Email: 778han@naver.com
8 Corresponding Author, Professor, Department of Liberal Arts, Korean Bible University, Republic of Korea, Email: hjinob@bible.ac.kr

passengers, and much research is being conducted on this. Researchers are proposing a system that automatically warns of danger through AI learning by using all data centered on the driver, such as face, body signals, driver appearance, gaze, voice, and surrounding sounds, as information for understanding the driver's state. The driver's facial information is considered the best information for judging the driver's emotions. Lee et al. (2018) proposed a CNN-based emotion classification method that distinguishes between aggressive and smooth driving using the driver's facial image. To this end, they detected the driver's emotions using near-infrared (NIR) and thermal imaging camera sensors and achieved 99% accuracy. Zaman et al. (2022) attempted to recognize seven emotions using a driver's face image. They used an improved Faster R-CNN and a neural architecture search network (NASNet), and their recognition accuracy was 99.15%. An attempt was also made to identify the driver's arousal state using measurements of body signals rather than facial expressions. Nakisa et al. (2020) used signals such as electroencephalography (EEG) and blood volume pulse (BVP) as CNN-LSTM learning data. They tested four driver states: high-arousal-positive emotions, low-arousal-positive emotions, high-arousal-negative emotions, and low-arousal-negative emotions. They showed an accuracy of 71%. Many studies also showed that learning a driver's appearance while driving could help predict safe driving by identifying distracted behavior. This is because drivers' increased use of digital devices such as mobile phones can decrease safety, regardless of their emotions. Ye et al. (2020) attempted to detect a driver's distracted behavior using a channel attention CNN method. This method used an SE module that assigns different weights according to importance. An experiment using the SE-Xception network showed an accuracy of 92.6%. Lyu et al. (2020) analyzed driver behavior by learning driver gaze and head pose information using CNN, and their analysis was accurate to 81%. Han & Cho (2021) trained 10 class data using CNN and improved ResNet-101 and achieved an F1 Score of 98.00%: Phone Use, Smoking, Music Adjustment, Drinking, Loss of Control, Looking Away, Navigation, Pet Care, Voice Call, and Smart Device Manipulation. Xu et al. (2022) showed an accuracy of 84% using an aggressive driving behavior prediction method using a Hidden Markov Model (HMM) and a Long Short-Term Memory (LSTM), and Park et al. (2024) proposed a method of learning driver abnormal behavior image data of six classes (Finding an item, Drunk driving, Drowsy driving, Vehicle control, Phone call, Mobile phone handling) using a Vision Transformer-based model to detect driver abnormal behavior, and showed an accuracy of 70-96% for each classification item.

Meanwhile, the results of studying human emotion recognition using voice information have been reported and show high accuracy. Choee et al.'s study (2019) used CNN-based transfer learning with English and German voices and a fine-tuning method to distinguish four emotions with 95% accuracy. Nam (2023) proposed DnCNN for voice emotion recognition exposed to noise. The average recognition rate of the voice without noise that distinguished five voices (neutral, happy, sad, angry, and fear) was 94.3%. When noise was added, it showed high recognition rates of 90.7% for subway transfer stations, 89.4% for cafeterias, and 90.4% for subways. Kim & Kwon (2023) developed an emotion recognition model using a multimodal method that simultaneously uses voice and image data. Voice data was represented as Mel Spectrogram, and image data was extracted using a Slowfast feature extractor. They showed that the model using 2D-CNN classified seven emotions and showed 60.11% accuracy. In this paper, we propose a multimodal driver status recognition system that uses images of the face, driving posture, and sound information, including the driver's voice. As interest in on-device AI increases, we tested whether it is possible to distinguish the status using images and sound information that can be measured with a regular mobile phone using a simple CNN structure. With the experiments, we found that the system was 99% accurate. The contributions of this paper are as follows.

(1) We proposed a multimodal CNN that learned the face and driving posture images, the driver's voice, and surrounding sound data to distinguish the driver's status. The experimental results showed 99.9% accuracy.
(2) We proposed a CNN system that distinguishes six conditions in total, including three driver conditions (normal, fainting, and drowsy) and three driver emotions (surprise, anger, and anxiety) that can cause abnormal driver states.
(3) Considering on-device AI, only images and sounds that can be measured with simple digital instruments were used, and it was shown that good results can be achieved even with simple CNN learning.

2. Related Works

Recent studies on driver status recognition for safe driving have focused on confirming various driver situations through images, measuring fatigue or stress levels by measuring body signals, and distinguishing between drowsy and non-drowsy drivers by analyzing the shape of the eyes and mouth on the driver's face. Each method shows good results, with an accuracy of 95% to 99%. In this study, we proposed a system that could distinguish three types of driver states: normal, faint, and drowsy, and three types of driver emotions that could cause abnormal driver states:

surprise, anger, and anxiety. This study combined existing studies recognizing driver states and emotions into one system. Its results are expected to be more useful in real life than existing studies. Table 1 shows examples of drivers' distracted behaviors used by ours and other researchers.

Table 1: Examples of drivers' distracted behaviors

Drivers' distracted behaviors	
Our six classification	c0: normal c1: faint c2: drowsy c3: surprise c4: anger c5: anxiety
Ten classification	c0: safe driving c1: texting on the phone - right hand c2: talking on the phone - right hand c3: texting on the phone - left hand c4: talking on the phone - left hand c5: operating the radio c6: drinking c7: reaching behind c8: hair and makeup c9: talking to passenger

Among recent papers proposing driver state recognition models, papers that studied drowsy driving classification have shown promising results. Kim et al. (2024) proposed a drowsiness recognition system based on SERN (Squeeze and Excitation Resnet Network) to recognize driver drowsiness. It learned with 99.84% accuracy, mainly from image data of the eyes and lips on the face. Das et al. (2024) used the CNN-LSTM algorithm with video clip and IoT sensor data to distinguish a driver's drowsy or non-drowsy state with 98.8% accuracy. In addition, other papers have captured the shape of the eyes and blinking on the driver's face and shown promising results. Madni et al. (2024) used CNN with the VGG16-LGBM method to capture image data of eye movement on the face and expressed the driver's drowsiness state with 99% accuracy. Ahmed et al. (2022) performed CNN-based learning, such as MTCNN and InceptionV3, with image data of the eyes and lips on the face and classified the drowsy driving state with 97.1% accuracy. Abbas and Alsheddy (2021) also used a method that combined CNN and RNN-LSTM models. They collected signals such as eye-blink images, electroencephalogram (EEG), heart rate (ECG), and electrooculogram (EOG) to measure driver fatigue, which can cause drowsy driving. Their method was accurate to 94.5%.

Meanwhile, some papers studied driver stress levels using body signals. Amin et al. (2023) classified driver stress levels into three categories (low, medium, and high) and used driver images and body signals such as ECG, HR, GSR, EMG, and RESP. Learning was performed by mixing CNN and LSTM methods, and stress levels were distinguished with 96.6% accuracy. Mou et al. (2021) used CNN and LSTM methods, along with images of the driver's eye condition and car and external environment data, to distinguish three stress levels with 95.5% accuracy. Doniec et al. (2020) used the driver's signal (EOG) information during actual driving to identify four driving activities (parking, roundabout, city traffic, and intersection). They used CNN to identify the driving status with 95.6% accuracy. Some papers distinguished drivers' distracted behaviors. Qin et al. (2022) implemented a small-sized D-HCNN using HOG (Histogram of Oriented Gradient) features extracted from the driver's image and distinguished 10 driver states with 99.87% accuracy. Huang et al. (2020) used a hybrid CNN using the driver's image and achieved 96.74% accuracy.

This study attempted to identify the driver's state more realistically by adding an emergency, such as fainting, to the simple distinction between drowsy and normal states previously presented. In addition, a model was proposed that distinguishes six driver states by distinguishing emotions such as surprise, anxiety, and anger that can cause abnormal driving conditions during driving through the driver's facial expressions. In addition to the driver's state image, the data used was expressed as sound data in spectrogram form. Considering the implementation of on-device AI, a simple CNN structure capable of multi-input in a multimodal manner was constructed, and an accuracy of 99.9% was achieved. Table 2 summarizes significant recent achievements and their research methods and compares them with the characteristics of this study.

3. Proposed System

We proposed a multimodal driver state recognition system, including a multi-input CNN that receives three image data inputs, learns through convolution layers, and goes through a concatenation process. As a factor considered in preparing the data, we collected images and sounds by utilizing cameras and sound recorders commonly installed in digital devices such as mobile phones, considering the trend of on-device AI. The driver's voice and environmental sound were imaged as spectrograms from sound wave files. The driver-face, driver-state, and sound spectrogram images were the three input data for learning and inference.

The proposed multi-input CNN structurally receives three image data, performs concatenation after two convolution layers, and then goes through a learning process with three convolution layers and three affine layers, and outputs six classes through the final softmax layer. This is a modified structure of the multi-input CNN proposed by Han (2024) based on Alexnet.

Table 2: Summary of significant recent achievements and their research methods of driver state

Authors	Number of class	Classification	Data	Multi-modal	Framework	Accuracy (%)
Ours	6	Six: drivers' distracted behaviors	Driver image, sound spectrogram	O	multi-input CNN	99.9
Kim et al. (2024)	2	Two: drowsy, Normal	Face image (eye, mouth)	X	CNN(SERN)	99.84
Das et al. (2024)	2	Two: drowsy, Normal	Driver image, signal (IoT sensors)	O	CNN-LSTM	98.8
Madni et al. (2024)	2	Two: drowsy, Normal	Face image (eye movement)	X	CNN (VGG16-LGBM)	99
Ahmed et al. (2022)	2	Two: drowsy, Normal	Face image (eye, mouth)	X	MTCNN, InceptionV3	97.1
Abbas & Alsheddy (2021)	2	Two: Fatigue, Normal	Eye-blink, signal (EEG,ECG,EOG)	O	CNN RNN-LSTM	94.5
Amin et al. (2023)	3	Three: stress levels (low,medium,high)	Driver image, signal (ECG, HR, etc.)	O	CNN-LSTM	96.6
Mou et al. (2021)	3	Three: stress levels (low,medium,high)	Driver image, Signal (eye, vehicle data, environmental data)	O	CNN-LSTM	95.5
Doniec et al. (2020)	4	Four: driving activity (parking, roundabout, city traffic, intersection)	Signal (EOG)	X	1D CNN	95.6
Qin et al. (2022)	10	Ten: drivers' distracted behaviors	Driver image (HOG feature)	X	D-HCNN	99.87
Huang et al. (2020)	10	Ten: drivers' distracted behaviors	Driver image	X	CNN(HCF)	96.74

Figure 1 shows the overall structure of the proposed multimodal driver state recognition system.

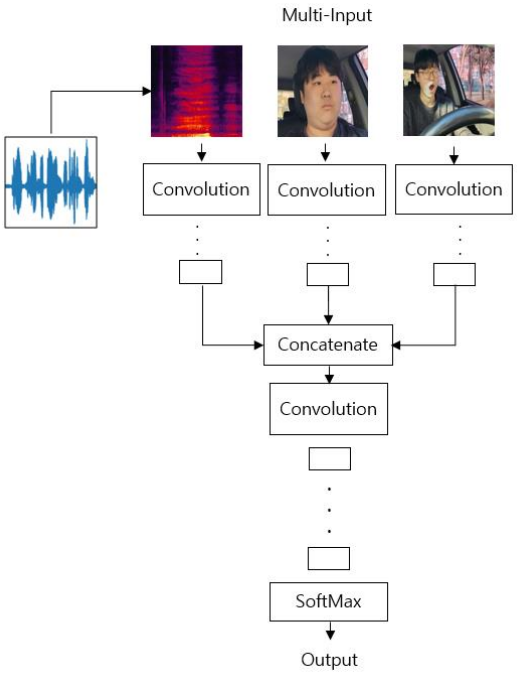

Figure 1: Overall structure of the proposed system.

Table 3 shows the detailed specifications of the eight layers multi-input CNN, including the parameters used for learning and inference.

Table 3: Detailed specifications of the multi-input CNN

Layer	Input 1	Input 2	Input 3
Input pixel size	227*227*3	227*227*3	227*227*3
First convolution	Filter 96 Channel 3 Width,Height 11 Stride 4 Zero padding 0 Output 55*55*96	Filter 96 Channel 3 Width,Height 11 Stride 4 Zero padding 0 Output 55*55*96	Filter 96 Channel 3 Width,Height 11 Stride 4 Zero padding 0 Output 55*55*96
First max pooling	Width,Height 3 Stride 2 Output 27*27*96	Width,Height 3 Stride 2 Output 27*27*96	Width,Height 3 Stride 2 Output 27*27*96
Second convolution	Filter 256 Channel 96 Width,Height 5 Stride 1 Zero padding 2 Output 27*27*256	Filter 256 Channel 96 Width,Height 5 Stride 1 Zero padding 2 Output 27*27*256	Filter 256 Channel 96 Width,Height 5 Stride 1 Zero padding 2 Output 27*27*256
Second max pooling	Width,Height 3 Stride 2 Output 13*13*256	Width,Height 3 Stride 2 Output 13*13*256	Width,Height 3 Stride 2 Output 13*13*256
Concatenate	13 * 13 * 256 * 3		
Third convolution	Filter 640, Channel 768, Width,Height 3, Stride 1, Zero padding 1, Output 13*13*640		
Fourth convolution	Filter 640, Channel 640, Width,Height 3, Stride 1, Zero padding 1, Output 13*13*640		
Fifth convolution	Filter 384, Channel 640, Width,Height 3, Stride 1, Zero padding 1, Output 13*13*384		
Fifth max pooling	Width,Height 3, Stride 2, Output 6*6*384		
First affine	Flatten 13824, Neuron 4096, Dropout 0.5, Output 4096		
Second affine	Neuron 4096, Dropout 0.5, Output 4096		
Third affine	Neuron 4096, Output 6		

4. Experimental Results

This experimental study hypothesizes that accuracy increases when multimodal data is used in a multi-input format for learning. In other words, we aim to prove that high accuracy values that cannot be achieved with single-input learning can be achieved with multimodal data in a multi-input format. Six class configurations were defined for the experiment, and data was created in the following manner.

c0: Normal is when the driver looks straight ahead, and the sound is expressed as relatively quiet ambient sound.

c1: Faint is when the driver looks down and rests his head on the steering wheel, and the sound is expressed as a cracking sound.

c2: Drowsy is when the driver closes his eyes or yawns, and the sound is expressed as a yawning sound.

c3: Surprise is when the driver opens eyes wide or mouth wide, expressing the sound as a horn sound.

c4: Anger is when the driver looks out the side window and gets angry, and the sound is expressed as an angry voice after the horn sound.

c5: Anxiety is the driver's tense expression, and the sound is expressed as a light horn sound, followed by an exclamation sound expressing anxiety.

The images used as the three input data were made to be the same size, 227 * 227 pixels. The sound spectrograms, which have a sound length of about 2 seconds, were created using Adobe Audition software. Figure 2 shows samples of the image data used in the experiment.

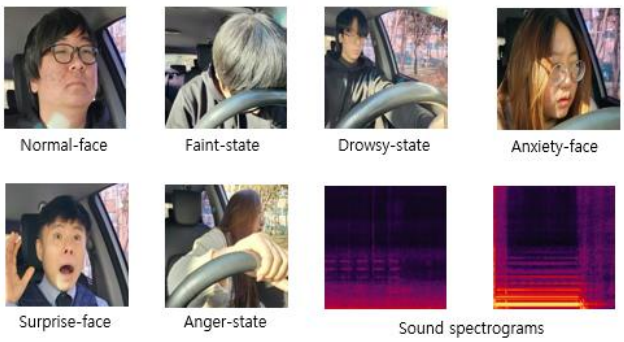

Figure 2: Examples of data

We created 18,900 images for the experiment, 12,600 of which were used as training data and 6,300 as test data. The learning rate was 0.001, and training was performed for 200 epochs. The other experimental specifications are shown in Table 4.

To evaluate the performance of Our Multi-input CNN used in the experiment, we conducted a parallel experiment comparing it with Resnet50 and Xception, which are well-known for their good performance. Xception consists of 14 modules and 36 convolution layers, all connected by residual connections except for the first and last convolution layers. ResNet50 has 50 layers, each trained by repeating different residual block shapes.

Table 4: Experimental specification

Item	Specification
Learning rate	0.001
Training epochs	200 epochs
Number of classes	6 classes
Number of data	Total data: 18,900 (6 classes * 3 inputs * 1050) Training data: 12,600 (6 * 3 * 700) Test data: 6,300 (6 * 3 * 350)
Data set	(1) Driver face images (2) Driver state images (3) Sound spectrograms

After training and testing using Resnet50, Xception, and our Multi-input CNN for single input data driver-face and driver-state, Resnet50 showed 81.0% and 85.7% accuracy, Xception showed 87.8% and 76.8%, and our Multi-input CNN showed 80.7% and 68.6% accuracy, respectively. Thus, all three models showed similar results. Therefore, Our Multi-input CNN can be considered suitable for learning prepared data. Meanwhile, the learning result of dual input: driver-face/ driver-state data was 75.8%, which was not much different from the single input, but the learning result of triple input: driver-face/ driver-state/ sound was 99.9%, showing that the learning of multimodal data using the multi-input method is effective. Table 5 shows the accuracy of the learning result according to the data set for each model.

Table 5: Experimental results

model	Data set	Accuracy (%)
Resnet50	Driver face Driver state	81.0 85.7
Xception	Driver face Driver state	87.8 76.8
Our Multi-input CNN	Driver face Driver state	80.7 68.6
	Dual input Driver face + state	75.8
	Triple input Driver face + state + sound	99.9

Figure 3 shows the train accuracy, test accuracy, and loss value for each of the four experiments on our multi-input CNN. In single-input learning for driver-face and driver-state, train and test accuracy showed similar trends during 200 learning epochs. However, the trend differed from

single input in the case of dual input, which simultaneously inputs the driver face and state and learns. The training accuracy showed an accuracy of over 99% from around 130 epochs, but the test accuracy stopped at 75% accuracy and did not increase any further. This phenomenon shows an overfitting problem in dual input image data learning. However, in the fourth triple input experiment, where sound spectrogram data was added, the test accuracy was confirmed to have improved to 99.9%. Therefore, our research hypothesis was proven that accuracy increases when multimodal data is trained as multi-input.

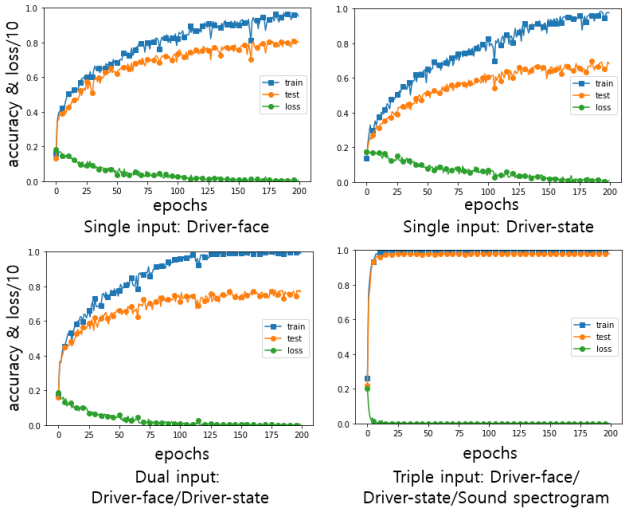

Figure 3: Test results of our multi-input CNN

5. Conclusions

This study attempted to prove the hypothesis that accuracy increases when multimodal data is trained in the form of multi-input through experiments. For the experiment, three types of data were prepared: driver-face, driver-state, and sound spectrogram. In single input training for driver-face and driver-state, the test accuracy was 80.7% and 68.6%, respectively. In the case of dual input training that simultaneously inputs driver-face and driver-state, the test accuracy was 75%, but in the triple input experiment where sound spectrogram data was added, the test accuracy was improved to 99.9%. Therefore, our research hypothesis that accuracy increases when multimodal data is trained in the form of multi-input was proven.

In this experiment, data was created assuming a digital device such as a mobile phone with a camera and recording function. We proposed a driver state recognition system with a simple multi-input CNN structure with a small number of layers that can be used for on-device AI. Before the era of 100% autonomous driving that does not require driver intervention arrives, driver responsibility still affects traffic accidents. In future research, if body signals such as electroencephalogram (EEG), heart rate (ECG), and electrooculogram (EOG) are collected using non-contact sensors and the collected data is utilized in a multimodal manner, it is expected that a system that can identify more diverse driver states can be proposed.

References

Abbas, Q., & Alsheddy, A. (2021). Driver Fatigue Detection Systems Using Multi-Sensors, Smartphone, and Cloud-Based Computing Platforms: A Comparative Analysis. *Sensors, 21(1), 56.* https://doi.org/10.3390/s21010056

Ahmed, M., Masood, S., Ahmad, M., & Abd El-Latif, A. A. (2022). Intelligent Driver Drowsiness Detection for Traffic Safety Based on Multi CNN Deep Model and Facial Subsampling. *in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 10, pp. 19743-19752, Oct. 2022.* https://doi.org/10.1109/TITS.2021.3134222

Amin, M., Ullah, K., Asif, M., Shah, H., Mehmood, A., & Khan, M. A. (2023). Real-World Driver Stress Recognition and Diagnosis Based on Multimodal Deep Learning and Fuzzy EDAS Approaches. *Diagnostics, 13(11), 1897.* https://doi.org/10.3390/diagnostics13111897

Choee, H. W., Park, S. M., & Sim, K. B. (2019). CNN-based Speech Emotion Recognition using Transfer Learning. *Journal of Korean Institute of Intelligent Systems Vol. 29, No. 5, October 2019, pp. 339-344.* http://dx.doi.org/10.5391/JKIIS.2019.29.5.339

Das, S., Pratihar, S., Pradhan, B., Jhaveri, R.H., & Benedetto, F. (2024). IoT-Assisted Automatic Driver Drowsiness Detection through Facial Movement Analysis Using Deep Learning and a U-Net-Based Architecture. *Information 2024, 15, 30.* https://doi.org/10.3390/info15010030

Doniec, R. J., Sieciński, S., Duraj, K. M., Piaseczna, N. J., Mocny-Pachońska, K., & Tkacz, E. J. (2020). Recognition of drivers' activity based on 1D convolutional neural network. *Electronics, 9(12), 2002.* https://doi.org/doi:10.3390/electronics9122002

Han, J. (2024). CNN-Based Multi-Factor Authentication System for Mobile Devices Using Faces and Passwords. *Appl. Sci. 2024, 14, 5019.* https://doi.org/10.3390/app14125019

Han, S., & Cho, J. (2021). Study on driver's distraction research trend and deep learning based behavior recognition model. *Journal of The Korea Society of Computer and Information, Vol. 26 No. 11, pp. 173-182, November 2021.* https://doi.org/10.9708/jksci.2021.26.11.173

Huang, C., Wang, X., Cao, J., Wang, S., & Zhang, Y. (2020). HCF: A Hybrid CNN Framework for Behavior Detection of Distracted Drivers. *in IEEE Access, vol. 8, pp. 109335-109349, 2020,* https://doi.org/10.1109/ACCESS.2020.3001159

Kim, J., & Kwon, J. (2023). Development of Emotion Recognition Model Using Audio-video Feature Extraction Multimodal Model. *Journal of the Institute of Convergence Signal Processing, Vol. 24, No. 4: 221~228 December 2023.* https://doi.org/10.23087/jkicsp.2023.24.4.007

Kim, M., Kim, W., & Choi, G. (2024). A Drowsiness Detection

System Using Multiple Driver Behavior Features Based on SERN. *The Journal of KINGComputing 2024 vol.20, no.3, pp.47 - 55.* http://doi.org/10.23019/kingpc.20.3.202406.005

Lee, K. W., Yoon, H. S., Song, J. M., & Park, K. R. (2018). Convolutional Neural Network-Based Classification of Driver's Emotion during Aggressive and Smooth Driving Using Multi-Modal Camera Sensors. *Sensors, 18(4), 957.* https://doi.org/10.3390/s18040957

Lyu, K., Wang, M., & Meng, L. (2020). Extract the Gaze Multi-dimensional Information Analysis Driver Behavior. *In Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20). Association for Computing Machinery, New York, NY, USA, 790–797.* https://doi.org/10.1145/3382507.3417972

Madni, H. A., Raza, A., Sehar, R., Thalji, N., & Abualigah, L. (2024). Novel Transfer Learning Approach for Driver Drowsiness Detection Using Eye Movement Behavior. *in IEEE Access, vol. 12, pp. 64765-64778, 2024,* https://doi.org/10.1109/ACCESS.2024.3392640

Mou, L., Zhou, C., Zhao, P., Nakisa, B., Rastgoo, M. N., Jain, R., & Gao, W. (2021). Driver stress detection via multimodal fusion using attention-based CNN-LSTM. *Expert Systems with Applications, Volume 173, 2021, 114693, ISSN 0957-4174.* https://doi.org/10.1016/j.eswa.2021.114693

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2020). Automatic Emotion Recognition Using Temporal Multimodal Deep Learning. *in IEEE Access, vol. 8, pp. 225463-225474, 2020,* https://doi.org/10.1109/ACCESS.2020.3027026

Nam, Y. (2023). Speech Emotion Recognition in Noisy Environments Based on a Denoising Convolutional Neural Network. *Journal of the Korea Institute of Information and Communication Engineering Vol. 27, No. 6: 772~781, Jun. 2023.* http://doi.org/10.6109/jkiice.2023.27.6.772

Park, M., Yoo, S., & Kim, J. (2024). A Deep Learning System for Driver Anomaly Behavior Classification using Vision Transformer. *Journal of Intelligence and Information Systems, 30(1), 59-74.* https://doi.org/10.13088/jiis.2024.30.1.059

Qin, B., Qian, J., Xin, Y., Liu, B., & Dong, Y. (2022). Distracted Driver Detection Based on a CNN With Decreasing Filter Size. *in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 6922-6933, July 2022.* https://doi.org/10.1109/TITS.2021.3063521

Xu, W., Wang, J., Fu, T., Gong, H., & Sobhani, A. (2022). Aggressive driving behavior prediction considering driver's intention based on multivariate temporal feature data. *Accident Analysis & Prevention, 164, 106477.* https://doi.org/10.1016/j.aap.2021.106477

Ye, L. et al. (2020). Using CNN and Channel Attention Mechanism to Identify Driver's Distracted Behavior. *In: Pan, Z., Cheok, A., Müller, W., & Zhang, M. (eds) Transactions on Edutainment XVI. Lecture Notes in Computer Science (TEDUTAIN, volume 11782), vol 11782. Springer, Berlin, Heidelberg.* https://doi.org/10.1007/978-3-662-61510-2_17

Zaman, K., Sun, Z., Shah, S. M., Shoaib, M., Pei, L., & Hussain, A. (2022). Driver Emotions Recognition Based on Improved Faster R-CNN and Neural Architectural Search Network. *Symmetry 2022, 14, 687.* https://doi.org/10.3390/sym14040687