Distracted Driver Posture and Head Position Identification via m-CNNs

1st Shyngyskhan Abilkassov
Nazarbayev University
Astana, Kazakhstan
shyngyskhan.abilkassov@nu.edu.kz

2nd Merey Kaiyrgaliyev
Nazarbayev University
Astana, Kazakhstan
merey.kaiyrgaliyev@nu.edu.kz

3rd Bauyrzhan Zhakanov
Nazarbayev University
Astana, Kazakhstan
bauyrzhan.zhakanov@nu.edu.kz

***Abstract*—Driver distraction is one of the major problems around the World. In order to reduce the number of deaths and terrific circumstances on the road, the paper aims to study the identification of driver's distraction via two assisted cameras applied from the side and front of the driver. The first camera from the side is used for the purpose to record the driver's posture position, while the second camera is used for recording the front position of the driver. This work evaluates performance of a multimodal convolutional neural networks approach. Two separately pretrained ResNet50 convolutional neural networks were concatenated to output final result. Results have shown that using individual CNNs performs better than a multimodal complex model with an accuracy of 71 %.**

***Index Terms*—CNN, ResNet50, m-CNNs, image classification**

I. Introduction

Even though global trends are heading towards fully autonomous vehicles, human drivers are still an essential part of road traffic. However, human drivers have their own flaws, one of which is a human factor - they often tend to be distracted by various activities when faced with monotonous task, such as driving. It is of special importance for Kazakhstan - the country with fatality rate of 24.2 deaths per 100,000 habitants [1]. In 2018 alone, there were 15,771 car accidents on Kazakhstani roads, which caused in total of 2,096 deaths and 20,455 injuries [2].

Unfortunately, since this problem of high distraction during driving depends on the nature and physiology of human-beings, there is no mass-market available solution to continuously and non-invasively and non-intrusively assist people during driving. Additionally, due to the gradual introduction of semi-autonomous vehicles to the market, identification of driver's fitness of control transition identification may play a huge role. In order to prevent accidents caused by distraction, the monitoring system has to be designed. In this paper we propose a solution to this issue based on action recognition system using multimodal deep learning model. The aim of the system is to detect the "distracted state" of the driver, if he/she is performing actions not related to the driving process itself. The system takes frames from two cameras as inputs, forwards it to neural network model and the model outputs the predicted class.

II. Related Work

Driver's distraction level estimation and classification during car driving is a crucial component which can enormously improve driver and passenger safety and reduce overall accident rates on roads. This can be done by identifying indicators of critical risk to be monitored. These factors include distraction from the road for some time, being "lost in thought", conversations, cell phone use, and drowsiness are all can be used to compute driver's total cognitive load on driving.

Previously, several ways of monitoring of driver fatigue and distraction activity using biometrical data were proposed. Several works proposing using electrocardiogram (ECG) and estimating heart rate variability [3] [4] and measurement based on EEG changes [5] and machine learning algorithms based on them [6] were proposed. Although these techniques showed adequate performance, they were mostly conducted in laboratory conditions and on limited amounts of data. Lack of real-world tests of these estimation techniques reveal crucial problems related to them - their invasiveness, immobility and requirement of special knowledge for proper installation. This clearly shows practicality of using deep learning vision based algorithms.

Furthermore, several driving studies have shown that a driver's allocation of visual attention away from the road is a critical indicator of accident risk [7]. Multiple research papers have shown that estimating driver distraction and fatigue levels by monitoring eye state during driving can be effective for this task [8] and even have shown robust results for people wearing glasses [9]. However, this technique for measuring fatigue levels may fail in low lighting conditions, when drivers have a greater risk of being drowsy. Therefore, additional monitoring technique should be introduced.

It has shown that identification of external objects, like mobile phones, coffee cups, as well as overall driver posture classification is a great technique for driver's overall distraction level estimation and is able to classify driver's drowsiness levels. One of the advantages of such systems is their ability to operate in low light conditions and identify behaviors of the robustly [10] given an available dataset.

One of the closest works in terms of approach to the problem is Lex Fridman's work of identification driver frustration identification [11]. However, this work is considering only the

side or the front of the driver. We have decided to combine both of them and collected the data from the front and side of the driver where images have generated simultaneously.

III. Dataset Design

In order to generate our own dataset, 2 RGB cameras were used: one capturing the frontal image of the driver from near the rearview mirror position and second monitoring the driver from side, above the car door of passenger's seat. Both cameras capture images with dimensions 1920x1080 pixels. Software part of the image recordings used rosbag package to capture frames at the rate of 5 frames per second. In total, 11 persons were invited to take part in data collection as drivers. Thus, each driver's data consisted of approximately 900 frames.

The "driver" supposed to sit on the chair with a steering wheel in front of it and perform 7 different actions within 3 minutes. Those actions, that eventually make up dataset's classes, included:

- Watching on the road;
- Looking left;
- Looking right;
- Looking at the rearview mirror;
- Checking instrument cluster;
- Interaction with center stack;
- Leaning slightly forward with head down (as if sleeping or texting on the phone);

Then, the dataset is cleaned of "dirty" data and sorted accordingly for passing images of the same moment from both cameras as inputs to front and side models. As for data augmentation, only random rotation is applied to the images. Eventually the dataset consists of **18892** images in total, that is partitioned into training, validation and testing subsets.

IV. Architecture and Proposed Method

ResNet50 from Microsoft Research group is a model for training very deep CNNs [12]. This model is an adaptation of its previous version VGG neural network. The input size of the image is 224x224, and the convolutional layers has 3x3 filters. The network is ended up with average pooling layer and 1000-class fully connected layer with softmax function. In our model training, ResNet50 has modified since we have used transfer learning approach for image classification since it helps to save time and model is not needed to be trained from scratch. By freezing all convolutional layers and using weights pre-trained from ImageNet, we removed last fully-connected layers and inserted our desired ones. Afterwards,

Fig. 1. Seven classes of driver activity

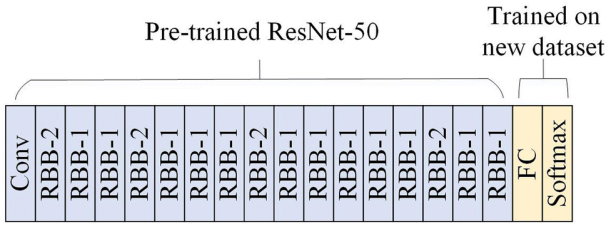

Fig. 2. Architecture overview of ResNet-50 for Transfer Learning

we have used ReLu activation function and added to each 512 long fully-connected layer, and finished with 7 long fully-connected layers.

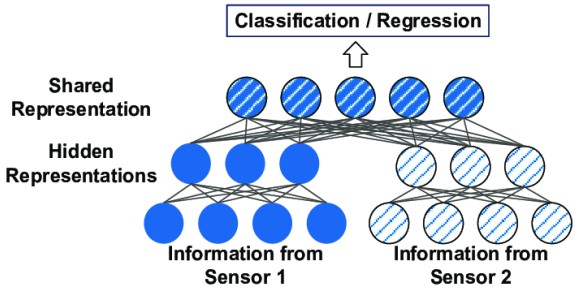

Fig. 3. Representation of multimodal deep learning approach layers

After training two individual ResNet50 networks for both posture classification using both frontal and side camera images, the multimodal deep learning approach was used. Firstly, last layers from both networks were cut off and each 512 long fully-connected layer was concatenated with each other finishing up in a 1024 long fully connected layer. Afterwards, after ReLu activation function additional 1024 long followed by 512 long and finally 7 long fully connected layers were added. ReLu activation was used after each layer. Result was obtained from last 7 nodes long fully connected layer.

V. Results

Initially, the separate models for front and side images were built and trained. As separate instances, both models showed satisfactory results, **80 %** (20 epochs) and **92 %** (24 epochs) for front image and side images classification respectively. However, when the multimodal deep learning approach was used for combining both of them, the overall accuracy has fallen to 71 % (50 epochs) against initial predictions of high accuracy due to "absorbing" best weights for each class from both models.

- Test Accuracy of Bottom: 87% (111/127)
- Test Accuracy of Center Stack: 68% (62/91)
- Test Accuracy of Instrument Cluster: 47% (56/119)
- Test Accuracy of Left: 66% (68/103)
- Test Accuracy of Rearview Mirror: 33% (41/123)
- Test Accuracy of Right: 46% (45/97)
- Test Accuracy of Road: 98% (323/329)

Test Accuracy (Overall): 71% (706/989)

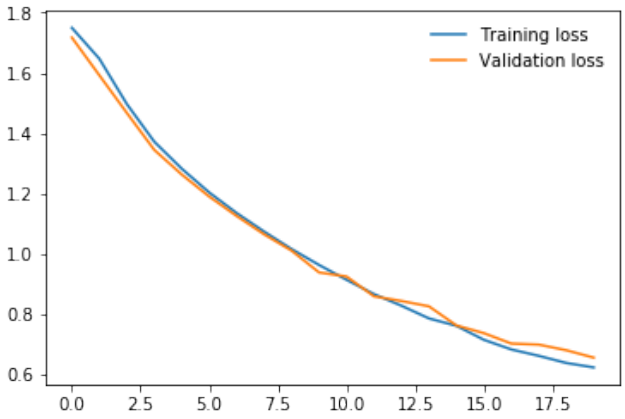

Fig. 4. Training progress for front image activity classification

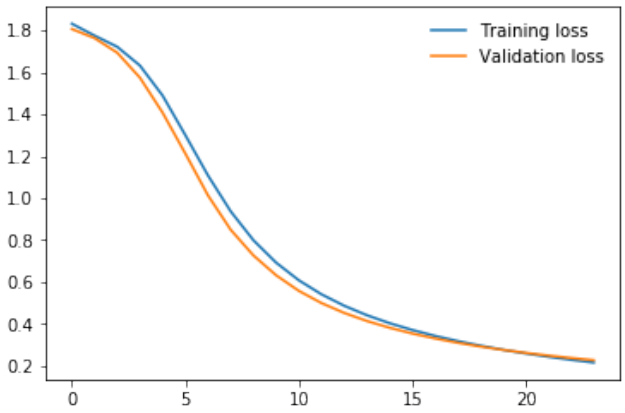

Fig. 5. Training progress for side image activity classification

The percentages above correspond to the individual class accuracy. Obviously, the ”Road” class has high accuracy level due to availability of more data. The ”Bottom” class is another one with high accuracy because this class' features are more distinct than others. Most other classes have medium accuracy, except for ”Rearview Mirror”, which has an accuracy of 33% because it is easily confused with ”Road” class.

Furthermore, one of the possible limitations of the work may occur due to fact that we have used too dense fully connected layers at the output. Simply connecting a 1024 fully connected contacatenated layer to the 7 neurons should may have performed better. Additionally, one should note the jumping behaviour of the validation loss curve. Authors assume that this strange behaviour may occur due to the way images were inputted to the final network. Images in the dataloaders were actually fed in a alphanumeric ordering without shuffling, so that same images from two datasets can be fed simultaneously to both ResNet50 networks. This means that same class and person images were fed to to the model by order, therefore network seems to be overfitting to one particular person at some instance which results in such jumps in validation loss and accuracy graphs.

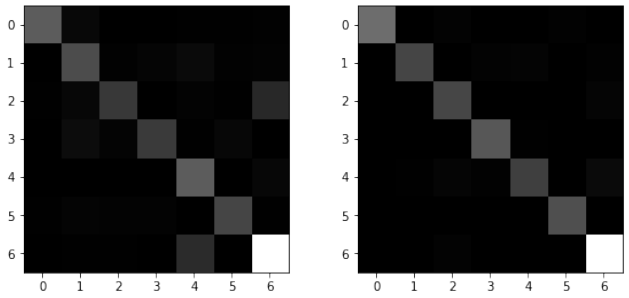

Fig. 6. Confusion matrices for frontal and posture image classification

Fig. 7. Examples of correctly classified frontal images

VI. Conclusion and Future Work

Driving on a road with high density traffic is an multi-agent activity that needs to consider an enormous number of factors. Until the automation of the whole world's traffic is reached, it will anyway lead to some casualties. However, the system proposed in this work is designed in order to prevent such casualties by detecting not related to the driving process actions. Overall, the obtained results can be considered as successful considering the difficulties faced during the development of the project. The overall accuracy could have been better without implementing multimodal deep learning approach, but adding complexity to the project encouraged the team to dive deeper into the theory and practice of building neural networks.

One the most feasible ways to improve accuracy and to achieve smooth validation loss curve may have achieved if new class of dataloader was implemented which can fed same instance images in a shuffled order. Additionally, at the current stage of project's development due to limitations in time and available resources, the initial plan was modified and adapted to circumstances. In the future, we would like to collect dataset from real car, and apply computer vision techniques for identification of driver head and posture positions to increase accuracy of the model and improve applicability of the system for real world conditions.

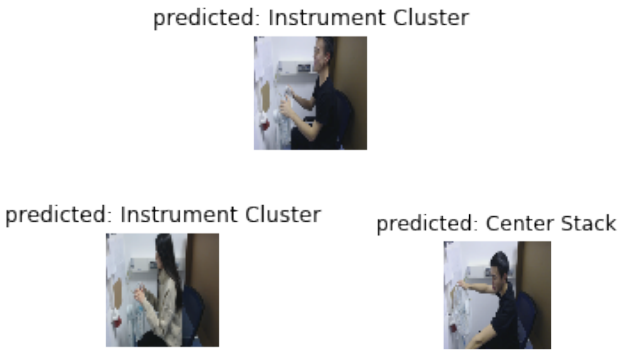

Fig. 8. Examples of correctly classified side images

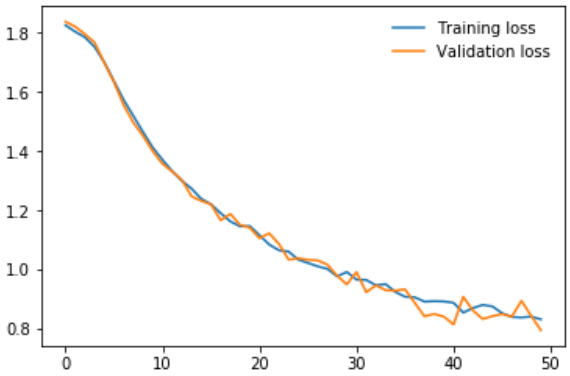

Fig. 9. Training progress for multimodal classification

ACKNOWLEDGMENT

We deeply want to say thanks to people who helped us to make this project. First of all, we want to thank our course supervisor professor Berdakh Abibullaev for his guidance throughout the course. We also want to thank professor Almas Shintemirov for providing space, time, and equipment for generating dataset and training the model in his laboratory. We have learnt so many things from Sergey Soltan, who also guided us throughout the project. Finally, we want to thank 8 people who had time and participated in our dataset generation.

REFERENCES

[1] World Health Organization (2018). Global Status Report on Road Safety 2018 (pp. 392–397).

[2] Committee on Legal Statistics and Special Accounts of the General Prosecutor's Office of the Republic of Kazakhstan (2018). Statistics of Road Traffic Accidents in 2013-2018. http://stat.gov.kz/official/industry/18/statistic/7

[3] Patel, M., Lal, S. K., Kavanagh, D., Rossiter, P. (2011). Applying neural network analysis on heart rate variability data to assess driver fatigue. Expert systems with Applications, 38(6), 7235-7242.

[4] Vicente, J., Laguna, P., Bartra, A., Bailón, R. (2016). Drowsiness detection using heart rate variability. Medical biological engineering computing, 54(6), 927-937.

[5] Jap, B. T., Lal, S., Fischer, P., Bekiaris, E. (2009). Using EEG spectral components to assess algorithms for detecting fatigue. Expert Systems with Applications, 36(2), 2352-2359.

[6] Mu, Z., Hu, J., Yin, J. (2017). Driving fatigue detecting based on EEG signals of forehead area. International Journal of Pattern Recognition and Artificial Intelligence, 31(05), 1750011.

[7] Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., Ramsey, D. J. (2006). The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data.

[8] Mandal, B., Li, L., Wang, G. S., Lin, J. (2016). Towards detection of bus driver fatigue based on robust visual analysis of eye state. IEEE Transactions on Intelligent Transportation Systems, 18(3), 545-557.

[9] Zhang, F., Su, J., Geng, L., Xiao, Z. (2017, February). Driver fatigue detection based on eye state recognition. In 2017 International Conference on Machine Vision and Information Technology (CMVIT) (pp. 105-110). IEEE.

[10] Yan, C., Coenen, F., Zhang, B. (2016). Driving posture recognition by convolutional neural networks. IET Computer Vision, 10(2), 103-114.

[11] Abdic, I., Fridman, L., McDuff, D., Marchi, E., Reimer, B., Schuller, B. (2016). Driver frustration detection from audio and video in the wild. Proceedings of the KI, 237.

[12] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).