Driver Drowsiness Detection System Based on Feature Representation Learning Using Various Deep Networks

Sanghyuk Park(✉), Fei Pan, Sunghun Kang, and Chang D. Yoo

School of Electrical Engineering, KAIST, Guseong-dong,
Yuseong-gu, Dajeon, Republic of Korea
{shine0624,feipan,sunghun.kang,cd_yoo}@kaist.ac.kr

Abstract. Statistics have shown that 20% of all road accidents are fatigue-related, and drowsy detection is a car safety algorithm that can alert a snoozing driver in hopes of preventing an accident. This paper proposes a deep architecture referred to as deep drowsiness detection (DDD) network for learning effective features and detecting drowsiness given a RGB input video of a driver. The DDD network consists of three deep networks for attaining global robustness to background and environmental variations and learning local facial movements and head gestures important for reliable detection. The outputs of the three networks are integrated and fed to a softmax classifier for drowsiness detection. Experimental results show that DDD achieves 73.06% detection accuracy on NTHU-drowsy driver detection benchmark dataset.

1 Introduction

Over the years, various safety-related driving assistant systems have been proposed to reduce the risk of car accidents, and statistics have shown fatigue to be a leading cause of car accidents. In fact, the American Automobile Association released a figure in 2010 that 17% of all fatal crashes in the USA could be attributed to tired drivers. This seems to be a global trend. In Germany, several studies conducted by Volkswagen AG in 2005 indicate that 5–25% of all collisions are caused by driver falling asleep. Failing concentration can impair steering behavior and reduce reaction time, and studies show that drowsiness increases the risk of collision by several factors. These figures indicate the need for a reliable driver drowsiness detection algorithm.

Drowsiness detection is studied by monitoring vehicle-based measurements, behavioral measurements, and physiological measurements. Vehicle-based measurements are from steering wheel movements, driving speed, brake patterns, and standard deviation of lane positions [8–12]. Behavioral measurements are obtained from driver eye/face movement using a camera. Physiological measurements such as heart rate, electrocardiogram (ECG) [7], electromyogram (EMG) [24], electroencephalogram (EEG) [6] and electrooculogram (EOG) [25] can be used to monitor drowsiness. Subjective measurements based on questionnaires and electrophysiological measures of sleep can also be made but it is

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 154–164, 2017.
DOI: 10.1007/978-3-319-54526-4_12

generally difficult to obtain drowsiness feedback from a driver in a real driving situation. There are reasons for and against using each measurement, and currently it is not clear which measurement is most reliable and cost effective. This paper is focused on detecting drowsiness by monitoring facial features and head movements of the driver.

It is assumed that drowsiness will manifest as rapid and constant blinking, nodding or head swinging, and frequent yawning [13,15–18]. PERCLOS (percentage of eyelid closure over the pupil over time) is considered a reliable measure for predicting drowsiness and has been incorporated in commercial products such as Seeing Machines and Lexus. Other facial movements such as inner brow rise, outer brow rise, lip stretch, jaw drop and eye blink have also been known to be markers for drowsiness. Until recently, most vision-based drowsiness detection has relied on hand-crafted features for monitoring facial and head movements. In general, hand-crafted features have shown limited effectiveness in real-world scenarios which might include drivers wearing sunglasses and large variation in illumination. On the other hand, features learned based on deep learning have been more effective in real-world scenarios. Choi *et al.* [14] developed a gaze zone detection algorithm based on features learnt using a convolutional neural network (CNN). Based on the learnt features, support vector machine (SVM) is used to predict drivers gaze zone. Dwivedi *et al.* [12] used a 3-layered CNN to learn facial features of detected face region from the input image. The outputs of the last layer are considered as the extracted features. On the basis of these features, the softmax classifier was trained and used for drowsiness prediction.

In this paper, we propose a deep architecture referred to as deep drowsiness detection (DDD) network in detecting driver drowsiness from input video. The proposed DDD network learns appropriate features for the task and predicts drowsiness of the driver. We consider both RGB video as well as optical flow as inputs. The proposed DDD network consists of three deep networks: AlexNet [1], VGG-FaceNet [2] and FlowImageNet [3]. Given an image sequences, the AlexNet is fine tunned to learn features related to drowsiness. The VGG-FaceNet is trained to learn facial feature related to drowsiness which is robust to genders, ethnicity, hair style and various accessories adornment. FlowImageNet takes dense optical flow image that is extracted from consecutive image sequences and is trained to learn behavior features related to drowsiness such as facial and head movements. Each three networks are independently fine-tunned for multi-class drowsiness classification given the following four classes: non-drowsiness, drowsiness with eye blinking, nodding and yawning. The softmax outputs of each networks are averaged for final classification. Before obtaining the final softmax output, fully connected (fc) 7 layer features of a block frames are concatenated for classification. We refer to this architecture as independently-averaged architecture (IAA). For comparison, the three networks are also integrated such that their fc7 layer features are concatenated, and based on this concatenated feature, input videos are classified into one of four classes. We refer to this architecture as feature-fused architecture (FFA). The proposed algorithm is evaluated on NTHU-driver drowsiness detection benchmark video dataset. The prediction results are presented in terms of detection accuracy. Experimental results show that DDD achieves 73.06% detection accuracy on NTHU-drowsy driver detection benchmark dataset.

The rest of this paper is organized as follows. Section 2 describes the proposed DDD network in detail. Section 3 presents experimental results comparing and analyzing various algorithms. Finally, Sect. 4 concludes the paper.

2 Proposed Deep Drowsiness Detection (DDD) Network

In this section, we present the proposed DDD network to detect driver drowsiness from complicated driving scenarios under varying circumstance. The proposed DDD network consists of two processes: learning feature representations and ensemble detection. During feature representation learning, we use three different networks: AlexNet, VGG-FaceNet and FlowImageNet. During drowsiness detection, we use two different ensemble strategies: independently-averaged architecture (IAA) and feature-fused architecture (FFA). The framework of DDD network for drowsiness detection using FFA is shown in Fig. 1.

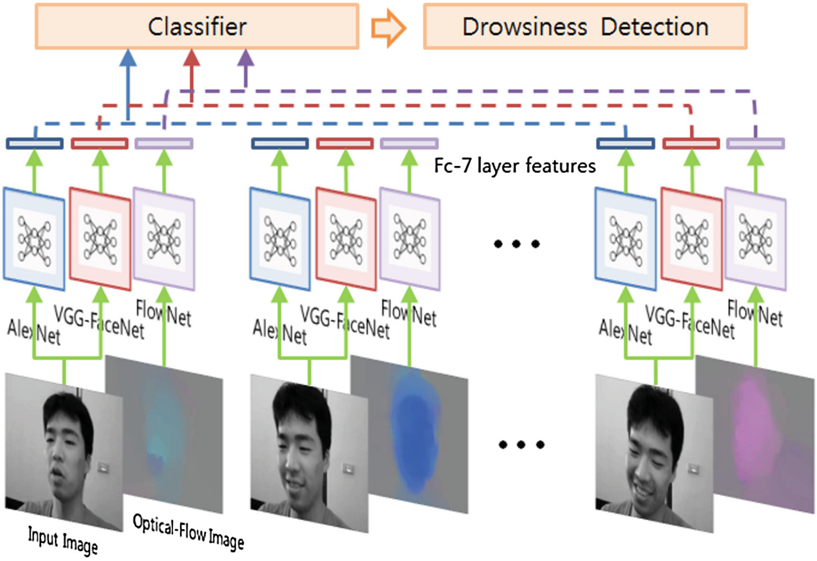

Fig. 1. The framework of deep drowsiness detection (DDD) network for drowsiness detection using feature-fused architecture (FFA).

2.1 Image Feature Representation Learning Based on AlexNet

To extract image feature which is robust to various backgrounds and environment changes (*i.e.*, indoor/outdoor, day/night) from the input image sequences, we adopt a pre-trained AlexNet model. A 8-layered AlexNet showed that deep

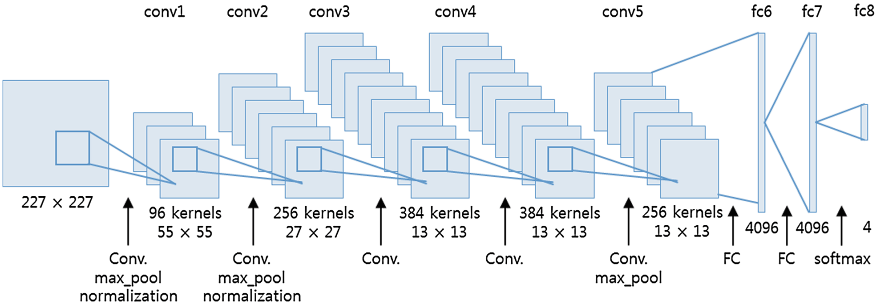

Fig. 2. The example architecture of AlexNet model for DDD network.

convolutional neural networks can significantly outperform other methods for the task of large scale image classification. This Alexnet consists of 5 convolution layers and 3 fc layers which has 60 million parameters and 650,000 neurons, and this model is trained with 1.2 million images for 1000 categories classification. For the proposed DDD networks, we fine-tunned AlexNet to classify multi-class drowsiness. The architecture and parameters of AlexNet model for DDD network are summarized in Fig. 2.

Each input image is down-sampled to size of 227 × 227. Training dataset is constructed by subtracting the mean value from pre-trained model, and AlexNet is fine-tunned using the drowsiness detection training dataset. The first convolutional layer filters the $227 \times 227 \times 3$ input image with 96 kernels of size $11 \times 11 \times 3$ and stride of 4 pixels. The second convolutional layer takes as input the (after normalized and pooled layers) output of the first convolutional layer and filters it with 256 kernels of size $5 \times 5 \times 48$. The third, fourth, and fifth convolutional layers are connected to one another without any intervening pooling or normalization layers. The third convolutional layer has 384 kernels of size $3 \times 3 \times 256$ connected (after normalized and pooled layers) to the outputs of the second convolutional layer. The fourth convolutional layer has 384 kernels of size $3 \times 3 \times 192$, and the fifth convolutional layer has 256 kernels of size $3 \times 3 \times 192$. The fully-connected layers have 4096 neurons. The output of the last fc layer is fed into a 4-way softmax layer which produces a distribution over the 4 class labels such as non-drowsy state, drowsy state with eye blinking, drowsy state with head nodding, and drowsy state with mouth yawning.

2.2 Facial Feature Representation Learning Based on VGG-FaceNet

To extract facial feature representation which is robust to facial characteristics (*i.e.,* genders, ethnicities) from the input image sequences, we adopt a pre-trained VGG-FaceNet model. A 16-layered VGG-FaceNet model was trained on various celebrity faces and evaluated on faces recognition task from the Labeled Faces in the Wild and YouTube faces datasets. The VGG-FaceNet consists of 13 convolution layers and 3 fc layers based on VGG-Very-Deep-16 CNN

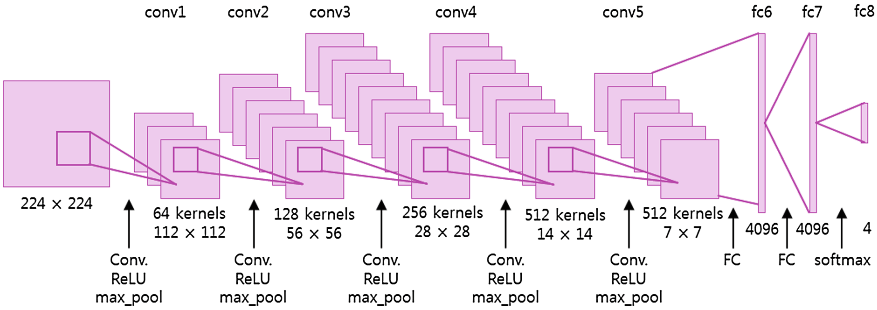

Fig. 3. The example architecture of VGG-FaceNet model for DDD network.

architecture, and this model is trained with 2.6 million images for 2,622 people recognition. The architecture and parameters of VGG-FaceNet model for DDD network are summarized in Fig. 3. Due to the deeper structure of VGG-FaceNet, we describe shorter version of VGG-FaceNet in Fig. 3. Similar to AlexNet training, we trained VGG-FaceNet with multi-class drowsiness classification about drowsy statements in fine-tuning manner. During the training, we down-sampled the images with a fixed resolution of 224 × 224, and the training dataset is constructed in the same way in AlexNet training. The training image dataset is passed through a stack of many convolutional layers, max pooling layers, Rectified Linear Unit (ReLU) activation function. Then these stacked layers are followed by three fc layers. The first two fc layers have 4096 neurons each, and the output of the last fc layer is fed into a 4-way softmax which produced a distribution over 4 class labels in the same way as training AlexNet.

2.3 Behavior Feature Representation Learning Based on FlowImageNet

To extract behavior feature representation which is related to movement patterns about drowsy states such as face and head gestures from the input image sequences, we adopt a pre-trained FlowImageNet model. A 8-layered FlowImageNet model consists of 5 convolution layers and 3 fc layers, and this model is used for video activity recognition task using UCF101 dataset [5]. The architecture and parameters of FlowImageNet model for DDD network are summarized in Fig. 4. As training AlexNet, FlowImageNet is trained using multi-class drowsiness classification about drowsy statements in fine-tuning manner. During training, we down-sampled the images with a fixed resolution of 227 × 227, and the training dataset is constructed in the same form as training AlexNet. Dense optical flow was calculated using [4] from consecutive image sequences and transformed into two channels of flow images by scaling and shifting x and y flow values to a range of $[-128, +128]$. A third channel for the flow image was created by calculating the magnitude of flows. The training image dataset is also passed through a stack of many convolutional layers, ReLU activation function,

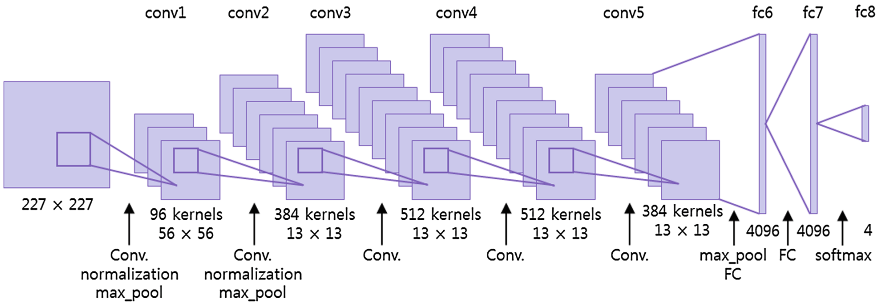

Fig. 4. The example architecture of FlowImageNet model for DDD network.

max pooling layers and normalization. A stack of various layers is followed by three fc layers. The first two fc layers have 4096 neurons each, and the output of the last fc layer is also fed into a 4-way softmax which produces a distribution over the 4 class labels in the same way as training AlexNet and VGG-FaceNet model.

2.4 Ensemble Detection Using DDD Network

The outputs of the three networks are combined to make single prediction. Ensemble model is well known to increase accuracy on various machine learning tasks [19–21], such as voting, averaging, stacking classifiers with regressors, and so on. For the proposed DDD network, we adopt two different fusion strategies: independently-averaged architecture (IAA) and feature-fused architecture (FFA). During IAA, the probability distributions of each network output for multi-class classification are integrated, and average probabilities are used to determine the driver drowsiness. During FFA, the three networks are also integrated such that their fc7 layer features are concatenated, and based on this concatenated feature, input video are classified into one of four classes using SVM [22].

3 Experiments

In this section, we provide competitive experimental results using proposed DDD network on drowsy driver detection video dataset. Due to the lack of previous benchmark performance on this dataset, we compare with the performance of several well-known classification algorithms such as variants of CNNs and LRCN [3].

3.1 Drowsy Driver Detection Video Dataset

To evaluate proposed DDD network, we use NTH Drowsy Driver Detection (NTHU-DDD) video dataset[1]. This video dataset contains 36 subjects including

[1] http://cv.cs.nthu.edu.tw/php/callforpaper/2016_ACCVworkshop/.

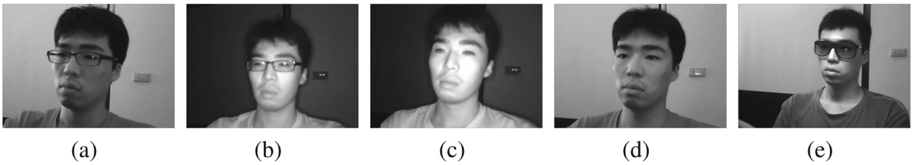

Fig. 5. Example frames of NTHU-DDD video dataset with different situations: (a) wearing glasses, (b) wearing glasses at night, (c) bare face at night, (d) bare face and (e) wearing sunglasses

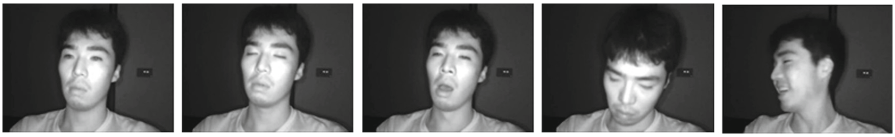

Fig. 6. Example frames of same situation (night bareface) and different behaviors (mixing drowsy and non-drowsy state) from 1 video clip.

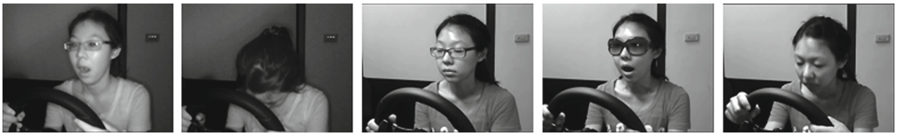

Fig. 7. Example frames of different situations (night wearing glasses, night bareface, wearing glasses, waering sunglasses, bareface) and same behavior (drowsy state) from 5 video clips.

different people, both genders, different ethnicities, which is in 5 situations as shown in Fig. 5. Each situation contains at least two behaviors about drowsy states, such as slow blinking, nodding, and yawning as shown in Fig. 6. The total dataset consists of train dataset, evaluation dataset, and test dataset. The train dataset consists of 360 video clips (722,223 frames) of 18 subjects. The evaluation dataset consists of 20 video clips (173,259 frames) of 4 subjects and test dataset consists of 70 clips (736,132 frames) of 14 subjects. During training and evaluation, each frame is binary labeled: drowsy or non-drowsy. The ground-truth label for test dataset is not publicly available yet. The dataset includes different physical attributes including variety in skin tone, fatigue, facial structure, clothes and hair styles. Thus, the algorithm for drowsiness detection should consider robustness and efficiency in all circumstances as shown in Fig. 7. The videos are in 640 × 480 pixels, 30 frames per second AVI format without audio.

3.2 Drowsiness Detection Experiments

The proposed DDD network aims to classify each frame in videos based on feature representation learning via different type of deep networks. Due to the lack of ground truth label of test dataset, we substituted evaluation dataset for test dataset. We train each convolutional neural network of DDD using 90% of train dataset, and validate proposed DDD network using the last 10% of train dataset. Firstly, train images are extracted 1 frame from every 10 frames in the videos. To improve robustness and generalization of our model, we adopted data augmentation based on pixel intensity. Each image was augmented by using intensity adjustment with 15%, 30%, −15%, −30%, and normalized into a range of $[0, 255]$. These training images resized to desired network input sizes are fed into each CNN. Further, training images are randomly fed into a multilayer CNN. The output of the last layer composed of 4 classes. All networks are trained using general fine tuning manner with modification of last fc8 and softmax layers. All weights are learnt via back-propagation and they are token as the learnt feature representations which are convolved with input images to produce the final feature representation to classify drowsy state. In our experiment, the output of the fc7 layer are considered as the trained feature representations. The feature representations are fed into the last softmax classifier and trained. Once the classifier has been trained, the whole images of evaluation dataset are tested on the trained classifier. We implement proposed DDD network using the MatConvNet package [23]. The batch size was set to 10, momentum to 0.9, and training was regularized by weight decay 10^{-4}. The learning rate was initially set to 10^{-3}, and then decreased by a factor of 10 when the validation set accuracy stopped improving.

Due to the lack of state-of-the-art performance on this dataset, we compared composed algorithm with the performance of several well-known multiclass classification algorithm such as variants of CNNs and LRCN. For this, we fine-tunned each AlexNet, VGG-FaceNet, FlowImageNet and LRCN using same training dataset and ground truth label, independently. For the human-level test results, all the extracted image sequences from the evaluation vdieo dataset were labeled manually as drowsiness state or non-drowsiness state by 5 human experts (doctoral students in computer vision area).

We evaluated the performance of individual CNN based models, LRCN, and proposed DDD network. The results are shown in Tables 1 and 2. In our experiments, DDD-IAA showed better results compare with DDD-FFA. For the DDD-IAA, we combine the outputs of three models of DDD by averaging their soft-max class probabilities. This improves the performance due to complementarity of the models. The average accuracies for different subjects were 70.81% (DDD-FFA) and 73.06% (DDD-IAA) which are higher than other CNNs including AlexNet, VGG-faceNet, FlowImageNet, LRCN as shown in Table 1. Because we ensembles the several features and sent into classification to yield outputs, which is robust to detect various drowsiness situation. Although these accuracies are lower than human experts's results but better than variants of previous CNN based models. Also, proposed DDD network showed better accuracies for different situation as shown in Table 2.

Table 1. Drowsiness detection accuracies for different subjects (%) on evaluation dataset

ID	Human	AlexNet [1]	VGG-FaceNet [2]	FlowImageNet [3]	LRCN [3]	DDD-FFA	DDD-IAA
004	73.37	61.12	66.20	65.92	52.65	78.26	66.87
022	83.42	80.30	77.80	53.36	77.31	78.64	86.27
026	85.07	56.68	66.12	59.33	58.86	68.16	69.00
030	81.47	65.62	61.28	67.40	63.15	58.16	70.11
Average	80.83	65.93	67.85	61.50	62.99	70.81	73.06

Table 2. Drowsiness detection accuracies for different situations (%) on evaluation dataset

Situations	Human	AlexNet [1]	VGG-FaceNet [2]	FlowImageNet [3]	LRCN [3]	DDD-FFA	DDD-IAA
Bareface	82.04	70.42	63.87	56.33	68.75	79.41	69.83
Glasses	78.83	61.63	70.53	61.61	61.73	74.10	75.93
Sunglasses	80.89	70.20	57.00	67.57	71.47	61.89	69.86
Night-bareface	82.54	64.69	73.75	66.82	57.39	70.27	74.93
Night-glasses	79.87	62.70	74.10	55.17	55.63	68.37	74.77
Average	80.83	65.93	67.85	61.50	62.99	70.81	73.06

4 Conclusion

This paper proposes a deep architecture referred to as deep drowsiness detection (DDD) network for learning effective features and detecting drowsiness given an input image of a driver. Previous approaches could only make decisions based on carefully hand-crafted features such as eye blinks and head gestures for detecting driver drowsiness. Deep network based feature representation learning approaches have been providing an automated and efficient set of learned features which help us to classify the driver as drowsy or non-drowsy very accurately. Especially, model ensemble are fusion strategies improves the performance due to complementarity of the models. Experimental results show that DDD achieves 73.06% detection accuracy on NTHU-drowsy driver detection benchmark dataset.

Acknowledgement. This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0101-16-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)) and supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2010-0028680).

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)

2. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC, vol. 1, p. 6 (2015)
3. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR, pp. 2625–2634 (2015)
4. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24673-2_3
5. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
6. Li, W., He, Q.C., Fan, X.M., Fei, Z.M.: Evaluation of driver fatigue on two channels of EEG data. Neurosci. Lett. **506**, 235–239 (2012)
7. Patel, M., Lal, S.K.L., Kavanagh, D., Rossiter, P.: Applying neural network analysis on heart rate variability data to assess driver fatigue. Expert Syst. Appl. **38**, 7235–7242 (2011)
8. Mattsson, K.: In vehicle prediction of truck driver sleepiness. Master's thesis, Luleå University of Technology, vol. 107 (2007)
9. Boyle, L.N., Tippin, J., Paul, A., Rizzo, M.: Driver performance in the moments surrounding a microsleep. Transp. Res. Part F: Traffic Psychol. Behav. **11**, 126–136 (2008)
10. Friedrichs, F., Yang, B.: Drowsiness monitoring by steering and lane data based features under real driving conditions. In: 2010 18th European Signal Processing Conference, pp. 209–213 (2010)
11. Forsman, P.M., Vila, B.J., Short, R.A., Mott, C.G., Dongen, H.P.: Efficient driver drowsiness detection at moderate levels of drowsiness. Accid. Anal. Prev. **50**, 341–350 (2013)
12. Dwivedi, K., Biswaranjan, K., Sethi, A.: Drowsy driver detection using representation learning. In: 2014 IEEE International Advance Computing Conference (IACC), pp. 995–999 (2014)
13. Lee, S.J., Jo, J., Jung, H.G., Park, K.R., Kim, J.: Real-time gaze estimator based on driver's head orientation for forward collision warning system. IEEE Trans. Intell. Transp. Syst. **12**, 254–267 (2011)
14. Choi, I.H., Hong, S.K., Kim, Y.G.: Real-time categorization of driver's gaze zone using the deep learning techniques. In: International Conference on Big Data and Smart Computing (BigComp), pp. 143–148 (2016)
15. Singh, M., Kaur, G.: Drowsy detection on eye blink duration using algorithm. Int. J. Emerg. Tech. Adv. Eng. **2**, 363–365 (2012)
16. Saito, H., Ishiwaka, T., Okabayashi, S.: Applications of driver's line of sight to automobiles-what can driver's eye tell. In: 1994 Proceedings of Vehicle Navigation and Information Systems Conference, pp. 21–26 (1994)
17. Horng, W.B., Chen, C.Y., Chang, Y., Fan, C.H.: Driver fatigue detection based on eye tracking and dynamk, template matching. In: 2004 IEEE International Conference on Networking, Sensing and Control, vol. 1, pp. 7–12 (2004)
18. Smith, P., Shah, M., da Vitoria Lobo, N.: Monitoring head/eye motion for driver alertness with one camera. In: ICPR, p. 4636 (2000)
19. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits Syst. Mag. **6**, 21–45 (2006)
20. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. **33**, 1–39 (2010)
21. Opitz, D., Maclin, R.: Popular ensemble methods: an empirical study. J. Artif. Intell. Res. **11**, 169–198 (1999)

22. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
23. Vedaldi, A., Lenc, K.: Matconvnet: convolutional neural networks for matlab. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 689–692 (2015)
24. Sahayadhas, A., Sundaraj, K., Murugappan, M.: Electromyogram signal based hypovigilance detection. Biomed. Res. **25**, 281–288 (2014)
25. Chieh, T.C., Mustafa, M.M., Hussain, A., Hendi, S.F., Majlis, B.Y.: Development of vehicle driver drowsiness detection system using electrooculogram (EOG). In: 1st International Conference on Computers, Communications, Signal Processing with Special Track on Biomedical Engineering, pp. 165–168 (2005)