Estimating Driver's Required Attention Level via Flow-Based Processing of Dash Cam Videos

Corcoran Gary, Clark James
Department of Electrical and Computer Engineering
McGill University
Montreal, Quebec
gary.corcoran@mail.mcgill.ca, clark@cim.mcgill.ca

***Abstract*—The problem addressed in this paper is to estimate the driver's perceived attention levels from in-car dash cam videos via flow-based processing. The input data consists of in-car dash cam videos coupled with a simple optical flow-based calculation between sampled pairs of frames. The overall model is built upon Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The dataset used is a collection of 1750 dash cam videos annotated as positive and negative examples of car accidents collected by Chan et al. [1]. Instead of using these binary accident categories, each video is newly annotated with a perceived attention level of the driver. The videos are collectively annotated among a group of subjects with the average score taken as the final attention label. The four categories of attention levels are as follows: low attention, medium attention, high attention, and very high attention. From these attention levels a deep neural network is used to estimate the driver's attentiveness in various situations. The complete system runs in real-time; approximately ten frames per second.**

***Keywords*-Deep Convolutional Neural Networks, Recurrent Neural Networks, Long-Short Term Memory Units, Driver's Attention, Dash Cam Videos, Active Safety Systems.**

I. Introduction

Even for the most experienced drivers, driving any vehicle is a very difficult task. This difficulty can be shown by the high number of collisions in Canada alone. In 2015, there were 118,404 collisions that were either fatal or involved a personal injury [2]. In about 84% of all accidents the cause was traced back to driver error [3].

Although the majority of vehicles are currently equipped with passive safety systems, i.e. systems to help reduce the outcome of an accident, such as seat belts, airbags, etc., there are still a high number of serious accidents. Newer intelligent car models are becoming equipped with active safety systems that utilize an understanding of the vehicle's state to avoid and minimize the effects of a crash. Some of these systems include collision warning and adaptive cruise control. Furthermore, research into these active safety systems have expanded into applications that work with or for the driver. This new generation of driver-assistance systems go beyond automated control systems by attempting to work in combination with a driver. These advanced safety systems include predicting driver intent [4], warning drivers of lane departures [5], etc.

Active systems have many benefits, however, they are difficult to implement as they require knowledge about the driver, the vehicle, or the environment.

To address this problem, we look into requiring knowledge about the driver's perceived attention levels. To do this, low-level flow fields from in-car dash cam videos are processed and a neural network is built to label the driver's required attention level into four categories: low attention, medium attention, high attention and very high attention.

In particular, the two components of our proposed system are as follows: 1) From input dash cam videos estimate the driver's perceived attention level in real-time, 2) When the perceived attention level of the driver reaches a category of high or very high provide an alert to the driver to help redirect their attention.

To build the system, simple flow fields are extracted from sampled pairs of frames captured by the in-car dash cam and are used as input into a Convolutional Neural Network (CNN). The CNN architecture used is an 18-layer Residual Network (ResNet) [6]. The last classification layer of the ResNet is removed and the resultant high-level flow features are used for further processing. These high-level features are extracted from each pair of frames and are passed into a Recurrent Neural Network (RNN). The RNN model uses Long Short-Term Memory (LSTM) units as building blocks for the network [7]. Since the proposed algorithm assigns a single attention level to the full video sequence, only the final hidden state of the RNN model is used. This last hidden state is then feed into a softmax classifier to produce the final output of the network, i.e. the driver's perceived attention level.

As unseen videos are passed through the network, the model breaks them into sequences of variable length. Each sequence is passed through the complete network and provided with a final attention label. If a given label is categorized as requiring high or very high attention, an alert is provided to the driver. The complete process runs in real time (ten frames per second) on a laptop GPU.

II. Related Work

With many new technologies becoming embedded into users' everyday life, drivers need to make sure they are paying adequate attention to their surroundings. To this end, there has been numerous works on modeling and monitoring driver's attentiveness. Many of this work attempts to directly correlate attention to another means of measurement such as drowsiness, head movements/position, or alertness [8], [9], [10], [11]. Although these systems can provide a broad understanding of the driver's attention level, there are always situations where a single measurement is not enough. In the case of driver drowsiness, relying solely on a measurement of how opened or closed a driver's eyes are would not be adequate in the event the driver was not drowsy but instead looking off into the distance for an extended period of time. In this case the driver's eyes would remain open, yet the driver would have a low attention level. In special cases like these additional measurements are required. For instance, measuring the driver's eye rotation could play an important role in detecting this anomaly. In general, many of these systems require multiple metrics to combat the numerous driving situations.

Another method currently used to monitor driver attentiveness is head tracking. Several researchers have worked on head tracking [12] to mixed success. Similarly, method [13] presents an approach which tracks the position of the head and estimates their respective head pose. It relies on 2-D template searching and a 3-D stereo matching.

Others systems attempt to use hardware solutions. [14] propose a system using infrared bean sensors to measure eye closure, and in turn, attention levels. This system works by placing infrared bean sensors above the eye to detect eyelid positions. When the eyelids interrupt the bean the systems will measure the time that the bean was blocked and thus providing eye closure measurements.

[15] propose a system using 3-D vision techniques to estimate and track the 3-D line of sight of a person. Their approach uses multiple cameras and multiple point light sources to estimate the line of sight without using user-dependent parameters.

Other systems [16], [17] rely on measuring external car behavior like the vehicle's current distance to roadway lines.

On the other end of the spectrum, some systems attempt to model the full cognitive architecture of human attention. These systems are designed with a single task such as car driving [3].

Many of these systems work well enough that newer vehicles are becoming equipped with these active driver monitoring systems, sometimes referred to as driver attention monitoring system. These vehicle safety systems were first introduced by Toyota in 2006 for its latest Lexus models [18]. These systems use infrared sensors to monitor driver attention levels. Specifically, these driver monitoring systems include a camera placed on the steering column which is capable of eye tracking via infrared LED detectors. In the case that the driver is not paying attention to the road ahead and a dangerous situation is detected, the system will warn the driver by flashing lights or providing warning sounds. If no action is taken by the driver, the vehicle will enter automation mode and apply the brakes (a warning alarm will sound followed by a brief automatic application of the braking system). After Toyota entered the market, various other companies began to follow their lead. Some examples of these are:

BMW offers an Active Driving Assistant with Attention that analyses the user's driving behavior and, if necessary, advises the driver to rest. The alert system to notify the driver as to when to take a break is provided in the form of graphic symbols shown on the control display [19].

Similarly, Bosch offers a driver drowsiness detection systems that takes input from a combination of the steering angle sensor, front-mounted lane assisting camera, vehicle speed, and turn signal [20]. Using this information a driver drowsiness level is computed.

Ford, Hyundai, and Kia all come fully installed with their respective driver attention warning systems. These were first debuted by Ford with their 2011 Ford Focus [21], Hyundai on their 2017 i30, and lastly Kia with their 2018 Stinger.

Mazda's driver attention alert activates at speeds above 65 km/h. This system learns driving behavior through steering input and vehicle road position during the beginning of the ride and compares the learned data to later stages. A difference above a set threshold triggers an audible and visual cue. This system was debuted on 2015 Mazda CX-5.

In 2009, Mercedes-Benz unveiled their safety system called Attention Assist which monitors the driver's fatigue and drowsiness levels based on their driving inputs [22]. It issues a visual and audible alarm to alert the driver if they are deemed too drowsy to continue driving. It is linked to the car's navigation system and can tell the driver where coffee and fuel are available.

Lastly, Nissan entered the market with their Driver Attention Alert (DAA), debuted with the 2014 Qashqai, followed by 2016 Maxima [23].

III. Our Approach

Given an input video sequence, the goal of our algorithm is to produce a single label corresponding to the perceived attention level of the driver. The input video consists of a sequence captured from a single dash cam. The dataset used is an accident dataset collected by Chan et al. [1]. The dataset consists of 620 dash cam videos captured in six major cities of Taiwan. From the 620 videos, 1750 clips were sampled where each clip consists of 100 frames (five seconds). These clips contain 620 clips where the moment of accident occurs at the last ten frames, and 1130 clips containing no accidents.

From this dataset, we labeled the driver's perceived attention levels. In order to do this labeling, each video was displayed to a user and the user manually annotated the video into one of the following categories: 1) Low attention, 2) Medium attention, 3) High attention, and 4) Very High attention. The labeling task was perform by three subjects who labeled all 1750 videos. From their assigned labels an average score, rounded to the closest integer value, is used for the final label. Figures 1 - 4 display sampled videos from each of the respective attention levels.

Initially our approach began by implementing the accident prediction model by Chan et al. [1]. To differentiate our approach from theirs, we wanted to achieve results in real-time. In order to attain this goal, it was necessary to locate and address the bottleneck in speed. In Chan et al.'s implementation, an advanced feature extraction process was used in which both appearance and motion features are computed. To capture the appearance, a fixed 4096 dimensional feature vector was extracted from each input video frame via a pre-trained VGG network [24]. For motion features, an improved dense trajectory feature [25] for a clip consisting of five consecutive frames was used. PCA helped reduce the dimensionality of trajectory features. Furthermore, a Gaussian-Mixture-Model was trained to produce a clustered feature vector. Lastly, a first-order statistic of fisher vector encoding was used to compute their final feature vector. For our implementation, it was required to remove a lot of this processing time to achieve a real-time implementation. To that end, it was decided to base our full implementation on simple motion estimations based on consecutive frames. The goal here is to see how much information can be gained from flow alone. From here, we wanted to move from the accident prediction problem to attention estimation using the same mindset. To do so, the labeled dataset was created and utilized.

The overall proposed algorithm is broken into three stages: 1) Motion Estimation, 2) Feature Extraction, 3) Attention Labeling. The following sections will describe each of these stages. The complete model is demonstrated in Figure 5.

A. Motion Estimation

The motion estimation used in our proposed algorithm is based on a dense optical flow field. This dense flow field computes optical flow for all points in a pair of frames. An implementation based on Gunner Farneback's algorithm is used [26].

To compute the flow field, a given input video sequence is sampled every five frames and motion estimates are computed between each pair of sampled frames. During training the video sequences are clipped to 100 frames and thus, the resultant flow field sequence length is 19. Additionally, during this stage each frame is resized to a dimensionality of 224×224. For the computation of optical flow, a window size, pyramid size, and polynomial degree of 15, 3, and 5, respectively, were used. This complete process resulted in a two-channel flow field with flow vectors in the x and y directions. These flow fields were normalized to a range of $[0-1]$. The final output of this stage is a sequence of flow fields with dimensions $19 \times 224 \times 224 \times 2$. Figure 6 demonstrates an optical flow output.

Table I
RESNET 18 MODEL

7×7, 64, stride 2
3×3 max pool, stride 2
$\begin{pmatrix} 3\times 3, 64 \\ 3\times 3, 64 \end{pmatrix} \times 2$
$\begin{pmatrix} 3\times 3, 128 \\ 3\times 3, 128 \end{pmatrix} \times 2$
$\begin{pmatrix} 3\times 3, 256 \\ 3\times 3, 256 \end{pmatrix} \times 2$
$\begin{pmatrix} 3\times 3, 512 \\ 3\times 3, 512 \end{pmatrix} \times 2$
average pool
1000-D FC

B. Convolutional Neural Network

The second stage of our algorithm is to extract higher-order features from the baseline optical flow features using a convolutional neural network. The CNN model used in our algorithm is a Deep Residual Network [6]. This ResNet is smaller than some of the newer models as it only consists 18 convolutional and two max-pooling layers. Table 1 summarizes the different layers in this network. Each flow field of dimensionality $224 \times 224 \times 2$ is treated as an input image and is passed through the CNN. Through convolution and pooling, the network extracts deep optical flow features that help incorporate additional information than that of the baseline flow fields. To extract these features, the last classification layer of the CNN model is removed, providing a feature dimension of 512. Once all flow sequences are passed through the CNN, the final output dimensionality is 19×512. The optical flow CNN features from each frame are then used as input into the next stage.

C. Attention Labeling

For attention labeling, a recurrent neural network is used. The RNN model used in this paper is a single-layer LSTM network consisting of 128 hidden units. Once all CNN flow features are computed on a given input video, each 512 dimensional feature is passed through the RNN sequentially. The RNN takes as input a given feature vector along with the previous resultant hidden unit. The output of the RNN model is a 128 dimensional feature vector at each time step in the sequence. Since a single attention label is required at the end of the video sequence, only the final 128 dimensional hidden

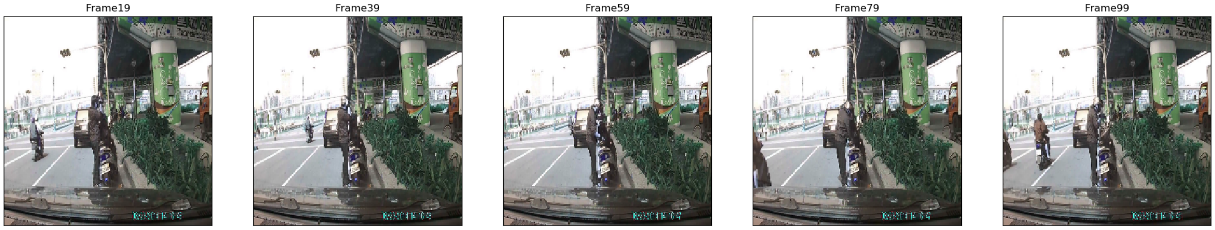

Figure 1. Low Attention Level Video Sequence

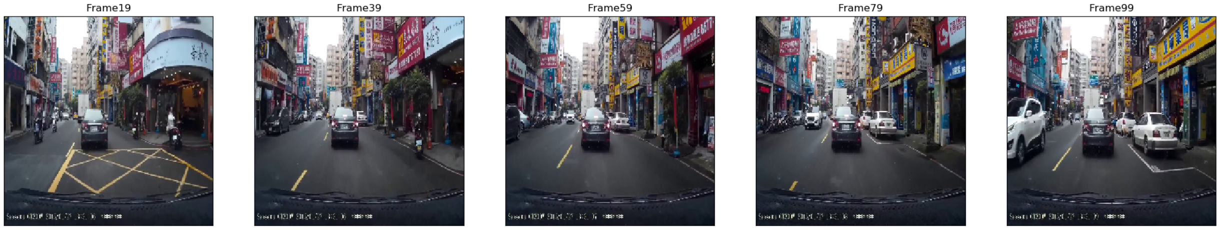

Figure 2. Medium Attention Level Video Sequence

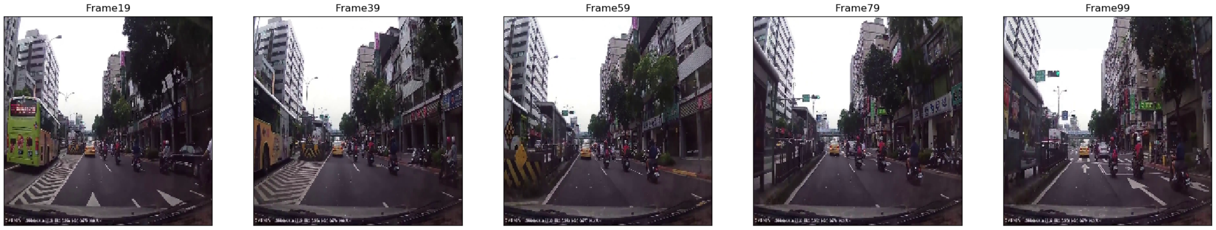

Figure 3. High Attention Level Video Sequence

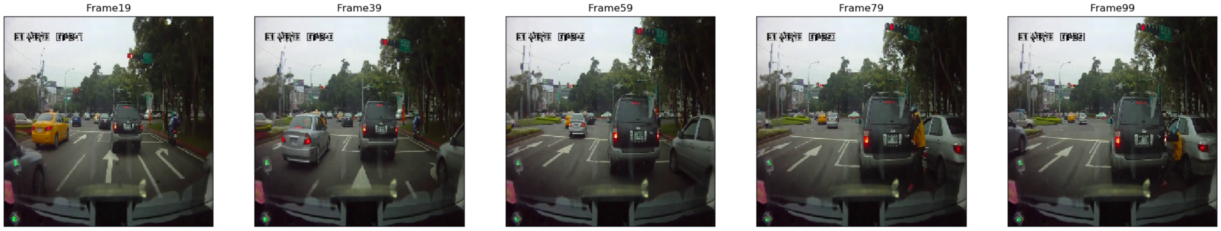

Figure 4. Very-High Attention Video Sequence

state is passed into the softmax layer. This layer reduces the resulting output size to the required number of labels.

D. *Implementation*

The complete attention estimation model is trained via back-propagation using a cross entropy loss function. Adam optimization [27] was used with a learning rate and batch size of 0.0001 and 50, respectively. Additionally, the CNN-RNN model is trained and tested with PyTorch [28] on a machine with eight cores, 16GB RAM, and a NVIDIA 970M GPU. The complete model takes approximately ten hours to train.

IV. Experimental Evaluation

In this section, we report our evaluation results on the driver attention dataset. We use a training, validation, and testing split of 60%, 20%, and 20%, respectively. This split results in 1050 videos used for training and 350 for validation and testing. Figure 7 demonstrates both the loss function and accuracy plots for both training and validation. As one can see from the graphs, our validation accuracy begins to plateau at approximately 50%. It is worth noting that the validation loss is lower than the training loss due to dropout layers. These dropout layers are placed between each pair of

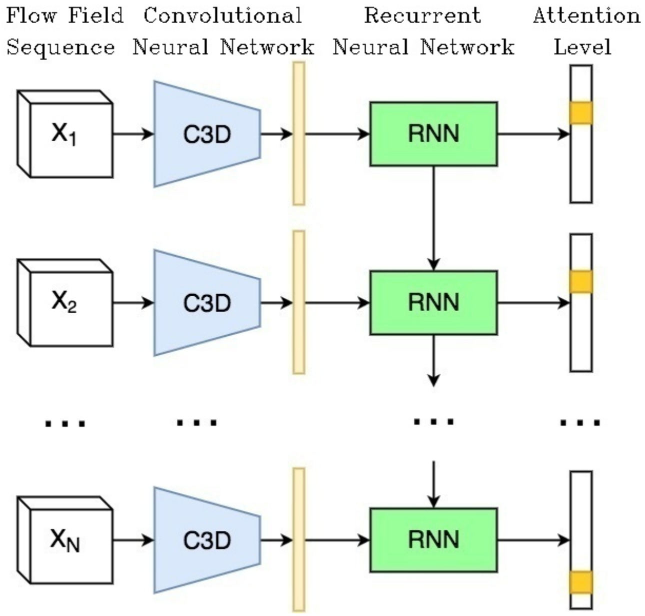

Figure 5. Network Architecture

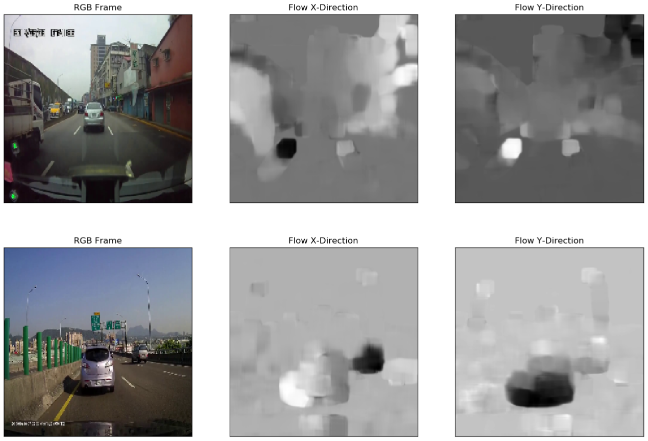

Figure 6. Optical Flow

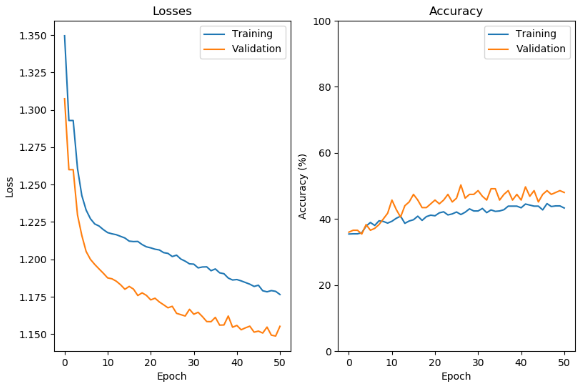

Figure 7. Training and Validation

Table II
MODEL ACCURACIES

Validation Accuracy (%)	Testing Accuracy (%)
50.29 %	50.74 %

convolutional layers. The behavior of this layer is different for training and validation cases. During training, a dropout rate of 50% is used. In the validation phase all features are used, and thus, the validation accuracy is more robust. This situation can lead to lower validation losses and higher validation accuracies, as seen in Figure 7. Table 2 displays the best validation score and the associated test score. The final test accuracy received was 50.74%. Since there are four categories, the complete algorithm produces a score twice that of random chance. The final result demonstrates that estimating the driver's attention level from flow-based processing is a tangible but difficult problem.

The confusion matrix in Figure 8 displays the validation accuracy among various classes. Each row signifies the correct label and each column the predicted label. The algorithm performs best on correctly classifying very high attention levels, however, it sometimes wrongly classifies them to high attention. Additionally, it has trouble classifying high attention situations and generally classifies them to either very high or medium levels.

It is worth noting that one of the features of our proposed system is to provide the driver with an alert when a high or very high attention level is required. To that end, when the overall system produces a false positive it is not necessarily unfavorable as any alert to help focus the driver would be beneficial. Additionally, a false negative is also not necessarily harmful as it produces the same result without driver aid. We believe that any accuracy over random chance provides additional safety to the end user.

Although the flow fields hold some knowledge that relates to the driver's attention level, it may not be enough information to provide a complete understanding of required attention levels. Future iterations of this work would attempt to include more features, namely appearance features, to help improve the accuracy without sacrificing the overall processing speed.

V. CONCLUSION

This paper presented the problem of estimating the driver's perceived attention via dash cam videos. Our proposed approach uses a low-level flow-based method coupled with a convolution and recurrent neural network. We achieve a real-time system that helps alert the driver when a high or very high attention level is required with an accuracy twice that of random change. Our final experimental results

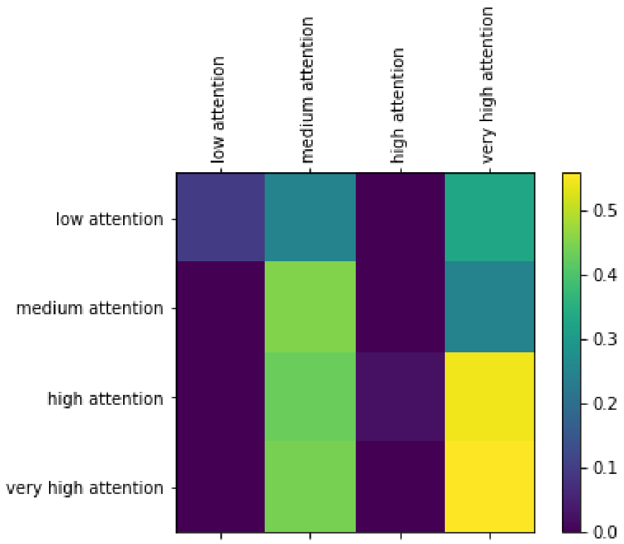

Figure 8. Validation Confusion Matrix

demonstrate the effectiveness of using simple flow-based features. Moving forward, we would like to continue this research to explore how including other fast processing features could help improve the accuracy while still maintaining results in real-time.

Acknowledgment

This work is supported by NSERC. Additionally, we gratefully acknowledge Brisk Synergies for providing funding for this research.

References

[1] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Computer Vision – ACCV 2016*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham: Springer International Publishing, 2017, pp. 136–153.

[2] "Canadian motor vehicle traffic collision statistics: 2015," https://www.tc.gc.ca/eng/motorvehiclesafety/tp-tp3322-2015-1487.html, 2017, accessed: 2018-02-08.

[3] K. Haring, M. Ragni, and L. Konieczny, "A cognitive model of drivers attention," in *Proceedings of the 11th International Conference on Cognitive Modeling, ICCM 2012*, 01 2012.

[4] D. D. Salvucci, "Inferring driver intent: A case study in lane-change detection," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 19, pp. 2228–2231, 2004. [Online]. Available: https://doi.org/10.1177/154193120404801905

[5] W. Kwon and S. Lee, "Performance evaluation of decision making strategies for an embedded lane departure warning system," *Journal of Robotic Systems*, vol. 19, no. 10, pp. 499–509, 2002. [Online]. Available: http://dx.doi.org/10.1002/rob.10056

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[8] Q. Ji and X. Yang, "Real time visual cues extraction for monitoring driver vigilance," in *Computer Vision Systems*, B. Schiele and G. Sagerer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 107–124.

[9] J.-D. Wu and T.-R. Chen, "Development of a drowsiness warning system based on the fuzzy logic images analysis," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1556 – 1561, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417407000401

[10] R. Onken, "Daisy, an adaptive, knowledge-based driver monitoring and warning system," in *Intelligent Vehicles '94 Symposium, Proceedings of the*, Oct 1994, pp. 544–549.

[11] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," *Sensors*, vol. 12, no. 12, pp. 16 937–16 953, 2012. [Online]. Available: http://www.mdpi.com/1424-8220/12/12/16937

[12] A. Gee and R. Cipolla, "Determining the gaze of faces in images," *Image and Vision Computing*, vol. 12, no. 10, pp. 639 – 647, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0262885694900396

[13] Y. Matsumoto and A. Zelinsky, "An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 499–504.

[14] T. Selker, A. Lockerd, and J. Martinez, "Eye-r, a glasses-mounted eye motion detection interface," in *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '01. New York, NY, USA: ACM, 2001, pp. 179–180. [Online]. Available: http://doi.acm.org/10.1145/634067.634176

[15] S.-W. Shih, Y.-T. Wu, and J. Liu, "A calibration-free gaze tracking technique," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 4, 2000, pp. 201–204 vol.4.

[16] A. Suzuki, N. Yasui, N. Nakano, and M. Kaneko, "Lane recognition system for guiding of autonomous vehicle," in *Proceedings of the Intelligent Vehicles '92 Symposium*, Jun 1992, pp. 196–201.

[17] D. J. King, G. Siegmund, and D. T. Montgomery, "Outfitting a freightliner tractor for measuring driver fatigue and vehicle kinematics during closed-track testing," in *SAE 942326*, 11 1994.

[18] "Toyota enhances pre-crash safety system with driver-monitoring function," https://newsroom.toyota.co.jp/en/detail/248128, 2015, accessed: 2018-02-08.

[19] "Bmw model upgrade measures taking effect from the summer of 2013," https://www.press.bmwgroup.com/global/article/detail/T0141144EN/bmw-model-upgrade-measures-taking-effect-from-the/summer-of-2013, 2013, accessed: 2018-02-08.

[20] R. B. GmbH, "Driver drowsiness," http://www.bosch-mobility-solutions.com/en/, 2013, accessed: 2018-02-08.

[21] "Driver drowsiness," https://web.archive.org/web/20110513232258, 2011, accessed: 2018-02-08.

[22] "Attention assist," http://media.daimler.com/marsMediaSite/en/instance/ko/Start.xhtml?oid=4836258, accessed: 2018-02-08.

[23] "2016 nissan maxima "4-door sports car" makes global debut at new york international auto show," http://nissannews.com/en-US/nissan/usa/releases/2016-nissan-maxima-4-door-sports-car-makes-global/debut-at-new-york-international-auto-show, 2015, accessed: 2018-02-08.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[25] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 3551–3558.

[26] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 363–370.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.