Vision-based system for monitoring vehicle operator responsiveness from face images

Quentin Massoz[1], quentin.massoz@ulg.ac.be [corresponding author]
Jacques G. Verly[1], jacques.verly@ulg.ac.be

[1] Laboratory for Signal and Image Exploitation (INTELSIG), Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium

Problem

Drowsiness is a major cause of fatal accidents, in particular in transportation. It is thus critical to develop real-time, automatic drowsiness monitoring systems designed to warn a vehicle operator (and/or an automatic driving system) to take timely safety actions. One of the least intrusive approaches for this is to use one or more cameras mounted in the vehicle, such as in the dashboard.

We report here on the preliminary results of an end-to-end system that takes grayscale video images of the operator's face (at 30 frames per second) as inputs, and produces an estimate of the probability density function (PDF) of the operator's reaction times (RTs) to sudden events.

Method

Our system is composed of two modules. The first produces, for each image in the video stream, the eyelids distance (i.e. the distance between the eyelids), and the second produces, from these distances over the last minute, the PDF of the RTs.

The first module consists of a convolutional neural network (CNN) taking as input a grayscale face image of size 128×128 and returning as output the eyelids distance, which is a positive scalar ranging from 0 to 6 pixels. The CNN is composed of ten 3×3 convolutional layers (interspersed with 2×2 max pooling layers), a global max pooling layer, and two fully connected layers. We use the Rectified Linear Unit (ReLU) non-linearity, Batch Normalization, and Dropout. In practice, the face input image is extracted from the full video frame using the Viola and Jones algorithm. Each frame is processed exactly once by this first module.

The second module is also a CNN. It takes as input a vector of size 1×1,800 that corresponds to the concatenated eyelids distances obtained over the last minute at the frame rate of 30 FPS, and returns as output two scalars: the log-mean and log-variance parameterizing a Gaussian PDF of the inverse of the RTs (=1/RTs). It is composed of a 1×15 convolutional layer, a 1×3 max pooling layer, a 1×31 convolutional layer, a global average pooling layer, and two fully connected layers. We use ReLU and Batch Normalization.

Results

For our preliminary results, we used 35 participants (21 females and 14 males) aged 23.3 ± 3.6 (mean ± SD) and free of drug, alcohol, and sleep disorders. The protocol - approved by the Ethics Committee of our university - led the participants to be in three successive states of increasing sleep deprivation of up to 30 hours, induced by acute, prolonged waking over two consecutive days. For each of these three states, each

participant performed a 10-minute psychomotor vigilance task (PVT), during which we recorded the RTs (in milliseconds) and the near-infrared face images with a Microsoft Kinect v2 sensor. However, due to some technical issues, only 88 out of the 105 PVT tests turned out to be usable.

The first module was trained using a Mean Squared Error (MSE) loss function. The sizes of the training set and validation set were 4,702 and 710 annotated face images, respectively. We obtained an MSE loss of 0.177 square pixels in training and of 0.234 square pixels in validation. Figure 1 illustrates some of the results, and figure 2 shows a scatter plot of the eyelids distance errors.

The second module was trained using a Negative Log Likelihood (NLL) loss function. We divided each 10-minute PVT into 37 one-minute segments, resulting in a training set and validation set of sizes 2,812 (31 subjects, 76 PVTs) and 444 (4 subjects, 12 PVTs) segments, respectively. Figure 3 shows a scatter plot of the predicted log-mean and log-variance.

Discussion

The method of the first module is slightly different from the typical face alignment method; here, it only returns the eyelids distance instead of multiple fiducial facial points, but this has the advantage of being fast and robust. Figure 2 shows that the first module achieves an error mostly below 1 pixel, and always below 2 pixels.

The second module successfully estimates the log-mean parameter, but has some difficulties with the estimation of the log-variance parameter (figure 3). The difficulty with the latter parameter is probably caused by two facts: (1) the low amount of RT observations within one minute (with an average of 9.4 RTs/min), resulting in a noisy target variance, and (2) the absence of discriminative features in the eyelids distance signal to estimate the variance.

One of the main novelties of our second module is the absence of pre-defined, hard-coded eyelids distance-based features, such as PERCLOS and mean blink duration. Instead, the learning algorithm automatically discovers informative features in the temporal signal, i.e., the features that are the most discriminative for predicting the PDF of the RTs. Overall, the only a priori assumption in our system is the design choice to focus on the eyelids. However, the automatic feature-discovery property makes the second module a “black box” that is hard to analyze and understand.

Both of the modules could be further improved by collecting, and training on, more data. Nevertheless, our results show that our two modules do not overfit their respective training sets, even with their modest size. Indeed, the overfitting is here mitigated by the relative small model size we chose (compared to other CNN applications with larger datasets) and a validation of the hyperparameters.

Summary

We developed and presented here the preliminary results of a novel, automatic, real-time, data-driven, end-to-end drowsiness monitoring system that feeds back the vehicle operator with the probability density function (PDF) of his/her reaction times to sudden events based on images of his/her face. In practice, the warned operator could use this

information, realize that there is a high (or higher than usual) probability that he/she will not be able to react fast enough to a sudden and potentially dangerous situation, and then decide - with full knowledge - to take safety actions such as pulling into the nearest rest area.

Future work includes collecting and training on more data, performing an in-depth analysis of the trained models, extending the system to more face modalities such as the head pose, and implementing different temporal model architectures.

Acknowledgements
Quentin Massoz is supported by a fellowship of the Belgian FRIA F.R.S.-FNRS. We thank Thomas Hoyoux, Philippe Latour and Clémentine François for the valuable discussions, and the NVIDIA Corporation for the donation of a GeForce GTX TITAN X.

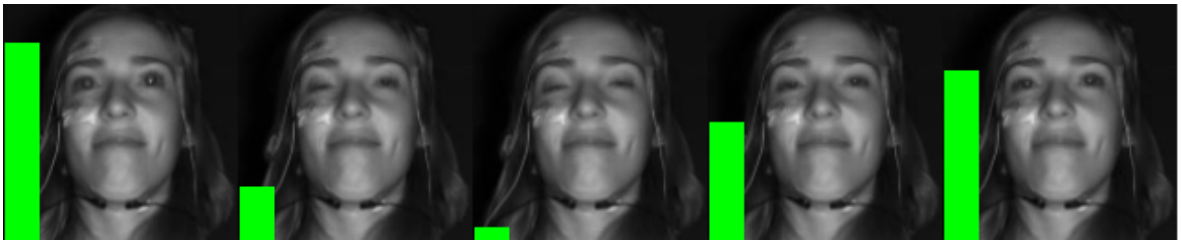

Figure 1: Five input/output results of the first module. The taller the green bar is, the larger the output - i.e. the eyelids distance - is.

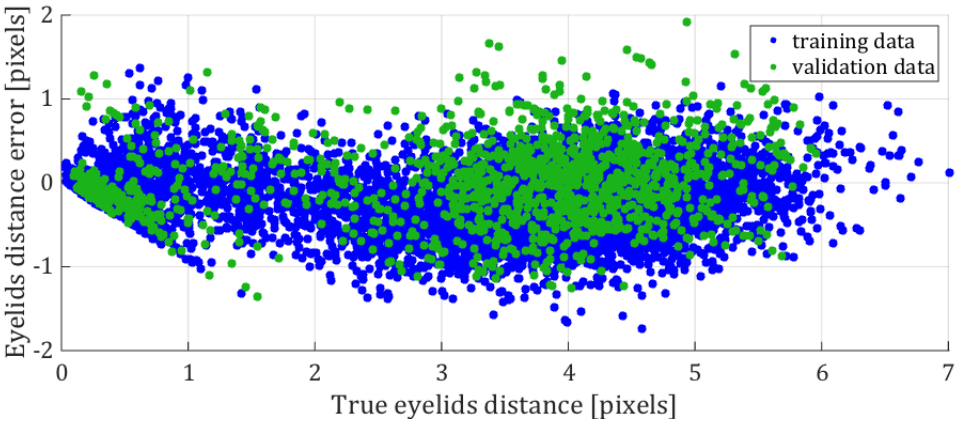

Figure 2: Scatter plot of the eyelids distance error versus the true eyelids distance.

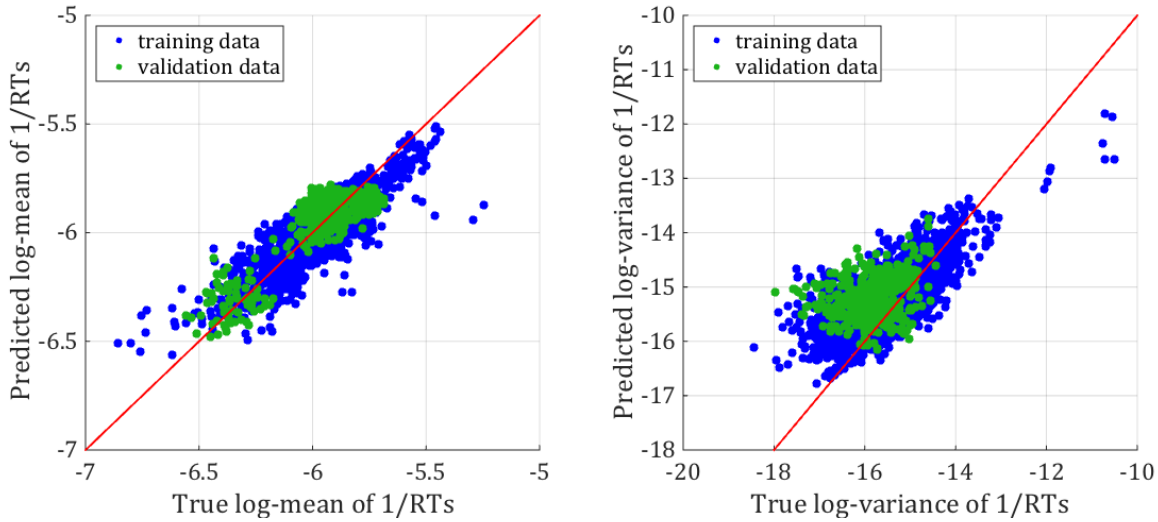

Figure 3: Scatter plot of the predicted log-mean (left) and log-variance (right) versus their respective true value.