A neural-network-based investigation of eye-related movements for accurate drowsiness estimation

Mingfei Sun[1], Masanori Tsujikawa[2], Yoshifumi Onishi[2], Xiaojuan Ma[1], Atsushi Nishino[3], Satoshi Hashimoto[3]

***Abstract*— Many studies reported that eye-related movements, e.g., blank stares, blinking and drooping eyelids, are highly indicative symptoms of drowsiness. However, few researchers have investigated the computational efficacy accounted for drowsiness estimation by these eye-related movements. This paper thus analyzes two typical eye-related movements, i.e., eyelid movements $X_{el}(t)$ and eyeball movements $X_{eb}(t)$, and investigates neural-network-based approaches to model temporal correlations. Specifically, we compare the effectiveness of three combinations of eye-related movements, i.e., $[X_{el}(t)]$, $[X_{eb}(t)]$, and $[X_{el}(t), X_{eb}(t)]$, for drowsiness estimation. Furthermore, we investigate the usefulness of two typical types of neural networks, i.e., CNN-Net and CNN-LSTM-Net, for better drowsiness modeling. The experimental results show that $[X_{el}(t), X_{eb}(t)]$ can achieve a better performance than $[X_{el}(t)]$ for short time drowsiness estimation while $[X_{eb}(t)]$ alone performs worse even than the baseline method (PERCLOS). In addition, we found that CNN-Net are more effective for accurate drowsiness level modeling than CNN-LSTM-Net.**

I. INTRODUCTION

Drowsiness has been studied for years in fatigue risk management and vigilance monitoring. Psychological studies have demonstrated that drowsiness can significantly impair productivity and the quality of work outcomes. For example, drowsy driving, usually caused by sleep loss, nights or very long working hours [1], is reported as one of the main causes of serious accidents [2], [3]. On the another hand, if drowsiness can be effectively estimated, the aforementioned side effects might be significantly mitigated or avoided. Studies show that accurate driver drowsiness estimation can help prevent potential accidents caused by driver fatigue [1]. Also, research on Massive Open Online Courses (MOOC) demonstrate that learning outcomes can be greatly improved by taking measures based on students' estimated level of drowsiness [4].

Eye-related movements, e.g., eyelid movements and eyeball movements, are reported to be important and useful for accurate drowsiness estimation by previous studies [5], [6], [1]. Some researchers used EOG techniques to explicitly collect eyeball movements for drowsiness estimation [5]. Another technique, the PERCLOS method [1], directly computes the percentage of eyelid closure over a short period of time as the drowsiness indicator, and is reported to deliver good results. Many drowsiness studies are thus focused on inferring eyelid movements [7], mostly via vision-based methods. Some used eye images/videos as they contain information of both eyeball movements and eyelid movements[8].

Despite the wide usages of eye-related movements in drowsiness estimation, few has tried to further differentiate the role played by different types of eye movements. Basically, there are two types of eye-related movements: eyelid movement $X_{el}(t)$ and eyeball movement $X_{eb}(t)$. The former is usually described by the degree of eye closure, including eyelid droops and blinks; while the latter often indicates gaze. Technically, investigation of these two types of eye movements is non-trivial since it requires the accurate extraction of eyelid and eyeball movements. Furthermore, although many neural-network-based temporal modelings, e.g., Convolutional Neural Network (CNN)[5] and Long Short-Term Memory (LSTM)[9], are proposed for modeling these eye-related movements, there is still a lack of insightful comparison studies on which model is better for drowsiness estimation.

In this paper, we conduct a systematic analysis of two types of eye-related movements for drowsiness estimation. The contributions of this paper are as follows. First, we analyze the computational efficacy of eyelid movements and eyeball movements for drowsiness estimation. Second, we propose and compare two neural networks for modeling temporal correlations of eye-related movements. Third, we conduct multiple experiments and present insightful interpretations as well as discussions for experimental results. Experiment evaluation shows that the combined eye-related movements are more effective than eyelid movements for short time drowsiness estimation, while eyeball movements alone fail to deliver good results. In addition, as drowsiness period extends, CNN-Net performs increasingly better than CNN-LSTM-Net for drowsiness level prediction.

II. METHOD

The proposed method includes five steps, as shown in Fig. 1. The details of each step are given as follows.

A. Facial Video Acquisition

The facial video data were collected from 29 subjects with their informed consent, and then labeled by 3 professional annotators. First, 29 subjects were asked to finish a specific math task sitting in front of a display with a web camera

This work was done when Mingfei Sun was interning in Data Science Research Laboratories, NEC Corporation. For collecting experimental data, the authors would like to thank both Yukihiro Kiuchi and Toshinobu Ogatsu belonging to Data Science Research Laboratories at NEC Corporation.

[1] Mingfei Sun (`mingfei.sun@ust.hk`) and Xiaojuan Ma (`mxj@cse.ust.hk`) are with Computer Science & Engineering Department, Hong Kong University of Science and Technology (HKUST).

[2] Masanori Tsujikawa and Yoshifumi Onishi are with Data Science Research Laboratories, NEC Corporation.

[3] Atsushi Nishino and Satoshi Hashimoto are with Technology and Innovation Center, DAIKIN Industries, LTD.

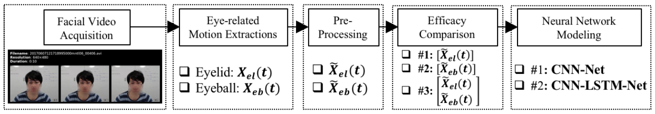


Fig. 1. The proposed investigation pipeline

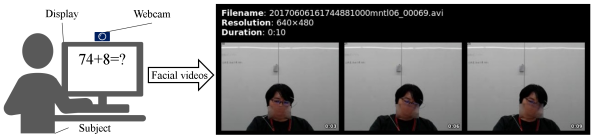


Fig. 2. Experiment setup for facial video collection

placed on top, as shown in Fig. 2. During the experiment, each subject was allowed to take a break, still remaining seated, for approximately half of the experiment time. The web camera captured the facial videos when subjects were performing the task or taking a break. All facial videos added up to an accumulated length of $205,600$s (~ 57h), with a resolution of $640 * 480$ and a frame rate of 30 fps. Second, 3 professional annotators rated every 10 seconds (without overlapping) of facial videos with 5 drowsiness levels as described in TABLE I [10]. For each 10 second video, the average of three labeled values is used as the ground truth of drowsiness level. The consistency of the labels is ensured by comparing annotations with the self-reported drowsiness status.

B. Motion Extraction & Pre-processing

The facial videos are used for extracting valid eyelid $X_{el}(t)$ and eyeball movements $X_{eb}(t)$, which are then pre-processed by a series of necessary manipulations.

Eye-related motion extraction: The $X_{el}(t)$ and $X_{eb}(t)$ are extracted at a frame rate of 30fps (the same as the video frame rate) by an accurate facial analysis tool developed by NEC corporation. To be more specific, $X_{el}(t) = \left[x_{el}^{L}(t), x_{el}^{R}(t)\right]$ where $x_{el}^{L}(t)/x_{el}^{R}(t) \in [0, 1]$ represent left/right eyelid openness, with 0 denoting fully open and 1 fully closed. Meanwhile, $X_{eb}(t) = \left[x_{eb}^{H}(t), x_{eb}^{V}(t)\right]$ where $x_{eb}^{H}(t)/x_{eb}^{V}(t) \in [-30, 30]$ denote horizontal/vertical degrees of gaze directions.

Pre-processing: Since the extracted $X_{el}(t)$ and $X_{eb}(t)$ are full of high frequency noises and redundancies, further filtering and down-sampling are required. Through multiple trials of different low-pass filters, the Hamming filter with window size 11 can effectively remove high frequency noises while also preserve the important peaks, e.g., blinks. Thus, we apply Hamming filters separately to $X_{el}(t)$ and $X_{eb}(t)$, and obtain the filtered signals ${X'}_{el}(t) = \left[{x'}_{el}^{L}(t), {x'}_{el}^{R}(t)\right]$ and ${X'}_{eb}(t) = \left[{x'}_{eb}^{H}(t), {x'}_{eb}^{V}(t)\right]$. Furthermore, in order to remove redundancies, ${X'}_{el}(t)$ and ${X'}_{eb}(t)$ are then half down-sampled to $\tilde{X}_{el}(t)$ and $\tilde{X}_{eb}(t)$, with 15 points per second. The examples in Fig. 3 illustrate this pre-processing.

TABLE I

DROWSINESS DESCRIPTIONS FOR DATASET ANNOTATIONS

Drowsiness level	Descriptions
1: Not drowsy at all	Fast and frequent gazes; stable eye blinks; active body moves
2: Slightly drowsy	Slow gazes; lips opening
3: Drowsy	Slow and frequent eye blinks; mouth moves; posture moves; face touches
4: Significantly drowsy	Frequent yawns; unnecessary body moves; slow blinks/gazes; deep breaths
5: Extremely drowsy	Eyelids closed; head tilting forward/backward

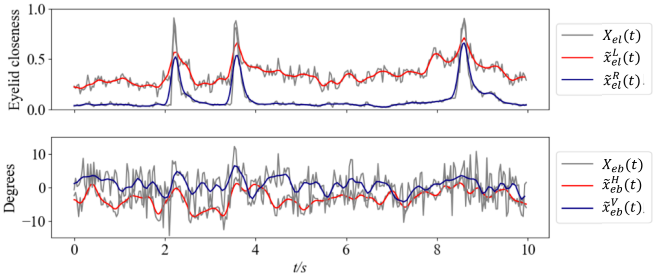


Fig. 3. Filtering & down-sampling of eyelid (top) and eyeball (bottom) movements

C. Efficacy Comparison & Neural Network Modeling

As for analyzing sequences of eyelid and eyeball movements, two typical types of neural networks can be employed, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), due to their reported better performances in many applications than conventional methods, e.g., Hidden Markov Model (HMM).

CNN [11], [5]: The CNN convolves input data (usually 1D or 2D), separately and stepwisely, with the same group of filters (fixed size) to automatically extract local patterns. Most of the filter parameters, except the filter number and size, can be learned and optimized according to sequence characteristics by training. Such advantage makes CNN a good model to extract temporal correlations for sequential data.

RNN [9]: Unlike CNN, RNN exploits its internal memory cells to track temporal correlations within an input sequence. However, due to the exploding and vanishing back-propagated errors, the conventional RNNs cannot capture long-term temporal correlations. The Long Short-Term Memory RNN (LSTM-RNN) has thus been proposed, which incorporates trainable forget gates to capture temporal flows along thousands or even millions of time steps. Similarly, all the memory parameters in LSTM-RNN, except the number of hidden variables, can be learned and optimized via sequence training. Owing to these advantages, LSTM-RNN is also a good model to spot temporal correlations from a time series.

To compare the efficacy of different eye-related movements for drowsiness estimation, we apply the same neural network structure, i.e., CNN-LSTM-Net as shown in Fig 4(a), to three different combinations of eyelid and eyeball movements: $\left[\tilde{X}_{el}(t)\right]$, $\left[\tilde{X}_{eb}(t)\right]$ and $\left[\tilde{X}_{el}(t), \tilde{X}_{eb}(t)\right]$. We use such architecture because of the complementarity of CNN and LSTM and its effectiveness for temporal signal

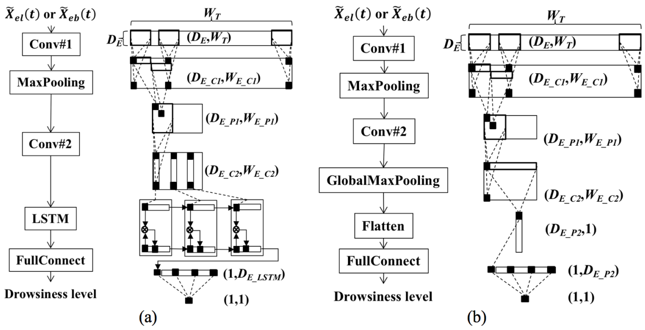


Fig. 4. Architectures: (a) CNN-LSTM-Net; (b) CNN-Net

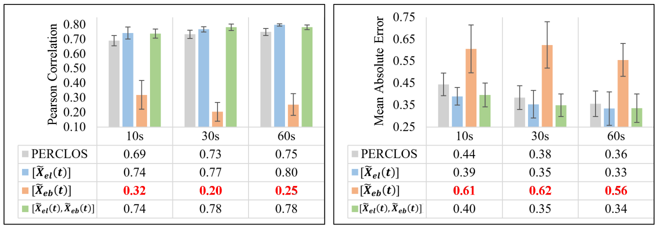
processing as reported in [12]. The CNN-LSTM-Net consists of two parts: CNN and LSTM. The former exploits two convolution layers and one max pooling layer to locally process every sequence segment, functioning as a feature extractor, while the latter models long-term correlations among these extracted features. For the joint movement input $\left[\tilde{X}_{el}(t), \tilde{X}_{eb}(t)\right]$, the $\tilde{X}_{el}(t)$ and $\tilde{X}_{eb}(t)$ are processed separately by the CNN part before being concatenated into a single vector sequence for the LSTM part.

Another neural network structure, i.e., CNN-Net as shown in Fig 4(b), can also be applied to estimate drowsiness from eye-related movements as reported in[5]. Different from CNN-LSTM-Net, CNN-Net contains only the convolution layers without the LSTM layers. We adopt a similar CNN structure as that in CNN-LSTM-Net, but replace the LSTM layer with a global pooling to obtain the local max values from segments.

III. EXPERIMENT RESULTS

The experiments are conducted on the facial video dataset, and two common regression metrics are used (Pearson Correlation and Mean Absolute Error) with the PERCLOS method [1] as the baseline. To remove the effects of individual-dependent mental correlations, the evaluation dataset is randomly split into five subsets (each with different groups of participants), and the cross validation is conducted by leave-one-group-out. Furthermore, in order to minimize the potential influences caused by the drowsiness durations, we extract facial video clips with different lengths (10s, 30s, 60s and 120s), and take the mean of all annotations as ground truth for the drowsiness levels.

In addition, the training parameters are determined based on training trials and the aforementioned cross-validations in a small subset of video data (test on the Nvidia 1080 8GB with the Adam optimizer). As a result, we set the batch size to 128 due to the fast convergence, and set the epoch to 100 for enough iterations. To further reduce the training time, an early stop checker was set up to monitor the validation loss and terminate the training process accordingly. The learning rate also shrinks according to the epochs for better training performance, starting from 0.001 and decaying by one order of magnitude after 20 and 50 epochs respectively.



	10s	30s	60s
PERCLOS	0.69	0.73	0.75
$[\tilde{X}_{el}(t)]$	0.74	0.77	0.80
$[\tilde{X}_{eb}(t)]$	**0.32**	**0.20**	**0.25**
$[\tilde{X}_{el}(t), \tilde{X}_{eb}(t)]$	0.74	0.78	0.78

	10s	30s	60s
PERCLOS	0.44	0.38	0.36
$[\tilde{X}_{el}(t)]$	0.39	0.35	0.33
$[\tilde{X}_{eb}(t)]$	**0.61**	**0.62**	**0.56**
$[\tilde{X}_{el}(t), \tilde{X}_{eb}(t)]$	0.40	0.35	0.34

Fig. 5. Results of using eyelid and eyeball movements to estimate drowsiness: Pearson Correlation (left) and Mean Absolute Error (right)

A. *Efficacy comparison*

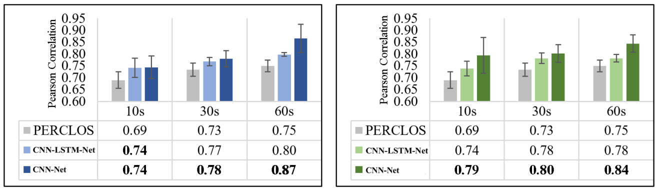
The efficacy of drowsiness estimation is compared among $\left[\tilde{X}_{el}(t)\right]$, $\left[\tilde{X}_{eb}(t)\right]$ and $\left[\tilde{X}_{el}(t), \tilde{X}_{eb}(t)\right]$. From Fig. 5, it is notable that using $\left[\tilde{X}_{eb}(t)\right]$ to estimate drowsiness fails to deliver good results, even worse than the baseline (PERCLOS): the Pearson Correlations for all video lengths indicate that the estimated drowsiness levels and ground truth labels are weakly correlated (Pearson Correlation< 0.4). Meanwhile, the Mean Absolute Errors (MAEs) for all video lengths are around 0.6, almost twice the MAEs achieved by the PERCLOS method. By contrast, the $\left[\tilde{X}_{el}(t)\right]$ and $\left[\tilde{X}_{el}(t), \tilde{X}_{eb}(t)\right]$ both perform better than the baseline and achieve superior results to $\left[\tilde{X}_{eb}(t)\right]$ for all three types of video lengths. Therefore, the eyeball movements alone are not an effective enough feature to estimate drowsiness while eyelid movements or joint movements are strongly related to drowsiness status.

The underlying reasons for why eyeball movements alone fail to capture the drowsiness may possibly be due to such phenomenon: when participants feel sleepy, their gaze will not move too much, and usually fix on one point. Consequently, the eyeball movements have few distinguishable differences to the situation when a participant is focusing their attentions with a sober mind. In the following experiments, the eyeball movements will not be considered.

B. *Neural network modeling*

The eyelid movements and the joint movements are input separately into two models, CNN-LSTM-Net and CNN-Net. As shown in Fig. 6(a), for eyelid movements, the CNN-Net deliver increasingly better results than that of CNN-LSTM-Net as video length increases. Similarly, considering the eyeball movements in Fig. 6(b), the CNN-Net consistently beats the CNN-LSTM-Net for all video lengths. Overall, the experimental results imply that CNN-Net modeling performs better (with stronger Pearson Correlations) than CNN-LSTM-Net modeling, and thus CNN-Net is more preferable for drowsiness estimation via eye-related movements.

We proposed one possible interpretation of why CNN-Net outperforms CNN-LSTM-Net for this case. The CNN-Net is naturally good at spotting short-time correlations due to its fixed-sized (usually small) filters. Also, by applying a same group of filters to each short segment, the CNN-Net "investigates" the short-time correlations from multiple perspectives. In contrast, the CNN-LSTM-Net is designed to automatically



(a)

(b)

Fig. 6. CNN-LSTM-Net vs. CNN-Net modelings for drowsiness estimation by using (a) eyelid movements $[\tilde{X}_{el}]$; (b) joint movements $[\tilde{X}_{el}, \tilde{X}_{eb}]$

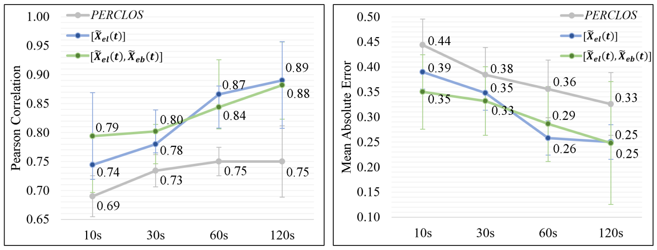


Fig. 7. Comparison of eyelid and joint movements via CNN-Net

"remember" information flows along many time steps and thus mainly captures the long-term temporal correlations. For drowsiness estimation via eye-related movements, most of the drowsiness status may possibly be indicated by instant body signals. For example, even though the eyelids are closed for a short time, longer than a normal blink, the person is still regarded as being sleepy. Such short-time eye movements appear to be more important and sensitive for drowsiness status than the long-term eye movements. Consequently, the short-term temporal modeling (by CNN-Net) is more likely to capture such important eye-related characteristics than long-term modeling (by CNN-LSTM-Net). Thus, the CNN-Net outperforms the CNN-LSTM-Net for eye-related drowsiness estimation.

To further confirm our analysis, we compare the eyelid movements and the joint movements for drowsiness estimation by CNN-Net modeling only. The results are shown in Fig 7. From the results, we find that the joint movements $[\tilde{X}_{el}, \tilde{X}_{eb}]$ deliver slightly better results for short time drowsiness estimation while eyelid movements alone achieve increasingly comparable or even better results as video length increases. Such performance boost indicates that the eyelid movements play an important role for drowsiness status estimation and CNN-Net can effectively capture these short-term correlations from the movements.

IV. Conclusion

In this paper, two typical eye-related movements (eyelid movements $\tilde{X}_{el}$ and eyeball movements $\tilde{X}_{eb}$) are analyzed for drowsiness estimation. Specifically, their computational efficacies are compared by inputing the associated features into a same model. In addition, two common neural network structures (CNN-Net and CNN-LSTM-Net) are investigated in terms of eye-related movement modeling. The evaluation results show that the eyelid movements alone $[\tilde{X}_{el}]$ as well as the joint movements $[\tilde{X}_{el}, \tilde{X}_{eb}]$ are effective indicators for accurate drowsiness estimation while eyeball movements alone $[\tilde{X}_{eb}]$ are weakly correlated with drowsiness status. Furthermore, the experimental results also demonstrate that the CNN-Net structure is more preferable to the CNN-LSTM-Net structure for eye-related movement modeling. We propose the possible interpretation that the short-term temporal correlations (modeled by CNN-Net) are more significant than the long-term temporal correlations (modeled by CNN-LSTM-Net) within eye-related movements, which is confirmed by further experiments. In addition, we also find that the joint movements can be more effective for short time drowsiness estimation while eyelid movements alone can still produce comparable estimations for long time drowsiness estimation.

References

[1] D. F. Dinges and R. Grace, "Perclos: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance," *US Department of Transportation, Federal Highway Administration, Publication Number FHWA-MCRT-98-006*, 1998.

[2] I. A. Akbar, A. M. Rumagit, M. Utsunomiya, T. Morie, and T. Igasaki, "Three drowsiness categories assessment by electroencephalogram in driving simulator environment," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 2904–2907.

[3] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, and J. Jang, "Real-time driver drowsiness detection for embedded system using model compression of deep neural networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 438–445.

[4] H. Y. Kim, B. Kim, J. Lee, and J. Kim, "Hey, wake up: Come along with the artificial learning companion to the e-learner's outcomes high!" in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2017, pp. 1763–1770.

[5] X. Zhu, W.-L. Zheng, B.-L. Lu, X. Chen, S. Chen, and C. Wang, "Eog-based drowsiness detection using convolutional neural networks." in *IJCNN*, 2014, pp. 128–134.

[6] W. W. Wierwille, S. Wreggit, C. Kirn, L. Ellsworth, and R. Fairbanks, "Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness. final report," Tech. Rep., 1994.

[7] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 818–830, 2014.

[8] L.-H. Du, W. Liu, W.-L. Zheng, and B.-L. Lu, "Detecting driving fatigue with multimodal deep learning," in *Neural Engineering (NER), 2017 8th International IEEE/EMBS Conference on*. IEEE, 2017, pp. 74–77.

[9] N. Zhang, W.-L. Zheng, W. Liu, and B.-L. Lu, "Continuous vigilance estimation using lstm neural networks," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 530–537.

[10] E. Zilberg, Z. M. Xu, D. Burton, M. Karrar, and S. Lal, "Methodology and initial analysis results for development of non-invasive and hybrid driver drowsiness detection systems," in *Wireless Broadband and Ultra Wideband Communications, 2007. AusWireless 2007. The 2nd International Conference on*. IEEE, 2007, pp. 16–16.

[11] S. Park, F. Pan, S. Kang, and C. D. Yoo, "Driver drowsiness detection system based on feature representation learning using various deep networks," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 154–164.

[12] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.