Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Detecting Behavioral Microsleeps using EEG and LSTM Recurrent Neural Networks

P. R. Davidson[1, 2], R. D. Jones[1, 2, 3, 4], M. T. R. Peiris[1, 2, 4]
[1]Van der Veer Institute for Parkinson's and Brain Research, Christchurch, New Zealand
[2]Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand
[3]Medicine, Christchurch School of Medical and Health Sciences, University of Otago, Christchurch, New Zealand
[4]Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand

Abstract—**Lapses in visuomotor performance are often associated with behavioral microsleep events. Experiencing a lapse of this type while performing an important task can have catastrophic consequences. A warning system capable of reliably detecting patterns in EEG occurring before or during a lapse has the potential to save many lives. We are developing a behavioral microsleep detection system which employs Long Short—Term Memory (LSTM) recurrent neural networks. To train and validate the system, EEG, facial video and tracking data were collected from 15 subjects performing a visuomotor tracking task for 2 1-hour sessions. This provided behavioral information on lapse events with good temporal resolution. We developed an automated behavior rating system and trained it to estimate the mean opinion of 3 human raters on the likelihood of a lapse. We then trained an LSTM neural network to estimate the output of the lapse rating system given only EEG spectral data. The detection system was designed to operate in real-time without calibration for individual subjects. Preliminary results show the system is not reliable enough for general use, but results from some tracking sessions encourage further investigation of the reported approach.**

I. Introduction

A short lapse in psychomotor performance at the wrong moment can have catastrophic consequences. Errors caused by fatigue while driving, for example, have been estimated to account for 10% of serious traffic accidents in France [1]. Fatigue has also been cited as a possible cause of the Exxon Valdez shipping disaster [2]. The early stages of fatigue are associated with gradual deterioration in perceptual, cognitive and sensorimotor performance [3, 4]. In deeper fatigue states it is common to observe sudden lapses of performance accompanied by behavioral features of sleep (head nodding, slow eye movements, loss of facial tone, partial or full eye closure [5]), followed rapidly by resumption of acceptable performance. These discrete events have been termed ‘lapses’, ‘microsleeps’, ‘blocks’ or ‘gaps’ [6], and have been shown to be correlated with changes in the EEG spectrum [7-9]. We will call these events *behavioral microsleeps*, or simply *lapses*, to emphasize the behavioral nature of the phenomenon.

While a relationship between EEG features and lapses has repeatedly been demonstrated [8-11], inter-subject reliability and temporal resolution are poor. Miller [12] noted that truck drivers were able stay within their lane while exhibiting up to 15 s of apparent EEG sleep. This is less surprising given that the relationship between the early stages of EEG and behavioral sleep is known to be weak [13].

Jung et al. [14] showed it is possible to use EEG spectra to predict minute-scale variations in lapse probability for an auditory task with a multilayer perceptron neural network (MLP). While the results were promising, inter-subject variability meant their detector needed to be individualized. Other recently developed systems have aimed to detect either drowsiness [15] or fatigue [16], as distinct from lapse events.

We are developing a system intended to detect and predict lapses from EEG data with second-scale temporal resolution. We intend the complete system to issue a warning indicating a lapse is imminent and trigger countermeasures to prevent further events occurring. At present the system employs data from a visuomotor tracking task study for training and validation.

Our system uses Long Short–Term Memory (LSTM) recurrent neural networks [20] to classify EEG feature vectors. LSTM networks overcome the "vanishing gradient" problem affecting most other recurrent neural network architectures when required to learn patterns over long time-lags, and have never previously been applied to lapse detection or, as far as we are aware, EEG analysis. In our system we train the neural network in a supervised manner using metrics derived from the behavioral data.

We have also developed a novel sub-system for identifying lapses based on behavioral data. The lapse identification system takes several behavioral metrics and uses an MLP network to derive an estimate of the probability of lapse once each second. The MLP network is trained on human rating data and can be considered to mimic human rating behavior. In this paper we report encouraging preliminary results for our lapse detection system.

II. Methodology

A. Tracking Study

In a previous study 15 normal male volunteers aged 18–36 years performed a visuomotor tracking task while we recorded EEG, video of facial features and tracking behavior (see [17] for full details). Approval for the study was obtained from a local ethics committee. Subjects were asked to keep a cursor as close as possible to a pseudo-random target (0.164 Hz bandwidth) scrolling down a screen (17” monitor, located 136 cm from the eyes) at 21.8 mm/s. The cursor was located at the bottom of the screen so subjects had an 8-s preview of the scrolling target. Subjects moved the cursor horizontally by rotating a steering wheel

0-7803-8740-6/05/$20.00 ©2005 IEEE.

(39.5 cm diameter; gain = 1.075 mm/deg). Angular position was sampled at 64 Hz. All subjects tracked the same target signal, facilitating inter-subject comparison. A 25 Hz analog video camera, time locked to the tracking, recorded head and facial features of subjects during the task. EEG and EOG data was recorded continuously during all sessions at 256 Hz. We recorded 16 channels of EEG from scalp electrodes placed according to the international 10-20 system, as well as horizontal and vertical EOG.

Each subject attended two sessions, held on separate days (mean inter-session interval 17 days, range 7–50), in which they performed the tracking task continuously for one hour. All sessions were held between 12.30 p.m. and 5.00 p.m. Cues such as time of day and remaining task time were not provided to subjects during the task. They were asked to stay alert and perform to the best of their ability and, aside from blinks, to keep their eyes open as much as possible during the task.

As part of the study, 30 hours of video were rated by a human expert (MP) who identified probable lapses, sleep, forced eye closure, distraction, and drowsiness with 1-s accuracy. The video analysis revealed that 8 of 15 subjects lapsed at some time during the two sessions. Of those that lapsed, the median rate was 44 lapses per hour. We also calculated a lower bound estimate of lapse frequency using the tracking data: lapses identified on video that coincide with a completely stationary cursor. The lower bound estimate gave a median count of 15 lapses per hour.

B. Lapse Rating Study

The video rating and lower bound estimates do not make optimal use of the behavioral data available. Ideally, we require a metric with good temporal resolution indicating overall lapse likelihood given all available behavioral data. Since opinion on when a lapse is occurring varies between experts, we conducted a short study to establish the mean opinion of human experts on the overall probability of lapse given both the tracking data and video data. Ideally, multiple human raters would view video of subjects' faces simultaneously with their tracking response and rate their moment-to-moment lapse likelihood. Previous experience showed this task would be extremely time consuming, amounting to month or more of work for each rater, and would need to be repeated on any new data collected. Consequently we had experts rate a subset of the data and built a "rating model" capturing their average opinion.

Three human experts were asked to rate 12 minutes of data from one randomly chosen session from each subject. The 12 minutes were composed of the first 2 minutes of each session followed by the period between 30 and 40 minutes from the start of the session. Raters were required to mark transitions between the levels of a 5-point scale capturing the degree of certainty that a lapse was occurring. The levels of the scale were labeled 1 to 5 corresponding to ratings of Definitely Not Lapse, Probably Not Lapse, Unsure, Probable Lapse, and Definite Lapse. Raters were required to make their judgment based on a combination of tracking behavior and the human video rating (previously carried out for the entire data set) displayed concurrently on a computer screen (Fig 1). We intend to further automate this process in future. The rating software showed a 30 s window of data which could be paged through or moved forwards and backwards in 1 s increments. The existing video rating was used as a proxy for metrics derived automatically and directly from the video data. We are currently developing these using computer vision techniques.

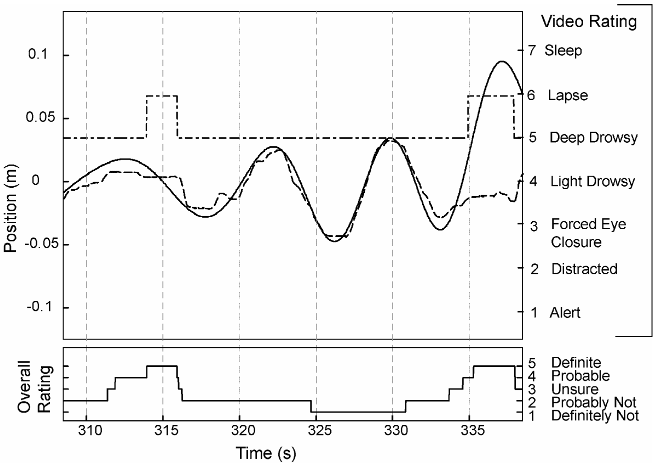

Fig 1. Interface used for lapse rating study. Raters were required to judge whether a person was lapsing given 30 s epochs of the target (solid line), the response (dashed line) and the opinion of an independent rater who was only shown video of the subject's face (dot-dashed line). Raters marked transitions between regions on a 5-point scale (lower part of display).

C. Automated Lapse Rating

We developed a lapse rating model which generated an estimate of the mean human rating given tracking and video data. In this task, lapses in tracking performance are most easily recognized as an extended period where the response cursor remains still while the target is moving (which we will call a *flat-spot*). Interpretation is confounded by the error-deadzone behavior characteristic of human tracking [18], which results in step-like tracking responses, particularly in low-velocity regions of the target. Other easily recognized lapse behavior is characterized by an erratic response incoherent with the target, leading to large tracking error. The lapse rating model comprised a MLP neural network with one 3-neuron hidden layer and 3 inputs: the human video rating, the absolute tracking error and the output of a flat-spot detector. The flat-spot detector activated whenever the cursor speed dropped below a fixed threshold ($5x10^{-5}$ m/s). Flat-spots shorter than 1.5 s were ignored and the output of the detector was scaled in proportion to the duration of the flat-spot. The 7-level video rating data was mapped from the integer values used by the human expert to the posterior probability the lapse rating was greater than 3 given only the video rating. These posterior probabilities were estimated using the same data set used to train the network.

Data from the lapse rating study was used to train the neural network. The training set comprised data from 7 randomly selected subjects and data from a further 8 subjects were put aside for validation. The MLP network was trained using Levenberg-Marquardt backpropagation (15 epochs, μ = 0.01), using the mean human rating as the target. After finding suitable MLP parameters, the network output was

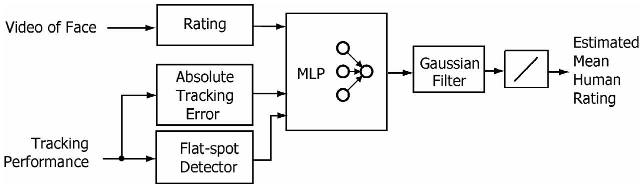

Fig 2. Rating model used to predict the mean expert rating based on tracking performance and video rating.

filtered using a bi-directional Gaussian filter (N = 60, α = 2.5). The filter output was then rescaled to fit the training target using linear regression.

D. Lapse Detection

EEG data from two posterior differential channels P3-O1 and P4-O2 were selected based on their good results in a preliminary investigation. Epochs exhibiting clear electrode pop were marked as artifact using a simple algorithm which detected a change of greater than 0.4 mV in EEG amplitude within a single sample. Both channels were then divided into sequential, non-overlapping 1-s windows. Power spectral density across each window was calculated using the covariance method to form a 40th order AR model. The covariance method was selected as it is known to work well for short data sequences [19]. The logarithm of the mean power in 7 frequency ranges was then calculated for each channel: delta ($0.1 < f \leq 4$ Hz), theta ($4 < f \leq 8$ Hz), alpha ($8 < f \leq 13$ Hz), low beta ($13 < f \leq 18$ Hz), high beta ($18 < f \leq 36$ Hz), gamma ($36 < f \leq 44$ Hz), higher ($f > 44$ Hz). These values were then converted into z-scores relative to the first minute of EEG data. This gave a 14-element feature vector for each second of EEG data which we used as input to our neural network model. The corresponding target was derived from the output of our rating model at the centre of the same 1-s time window. A binary threshold of 3 was applied to the rating model output, which became 1 if the model indicated a lapse and 0 otherwise.

Our system employed a Long Short-Term Memory (LSTM) dynamic neural network with forget gates [20]. We used a slightly modified version of the implementation included with PDP++ [21]. LSTM can learn patterns occurring in different time scales without specifying those timescales. In this paper we describe results from a network containing 6 LSTM blocks with 3 memory cells per block. A linear bypass between input and output was also provided. The network was trained according to [20], except that the output error was scaled in proportion to the confidence of the rating model. Confidence was estimated by finding the distance of the rating model output from 3 (the level corresponding to "not sure"). We used sequential online training, except that whenever a sample was identified as containing EEG artifact the weights were not updated and the internal states of all memory cells were reset. The training set was presented to the network 150 times, and we used learning rate $\mu = 1\times10^{-5}$ and momentum 0.9.

8 of the 15 subjects in the study lapsed at least once and only their data was used to train and test the network. To measure performance, we converted the continuous network output to a binary variable by applying a decision threshold. We selected the threshold giving maximum agreement between the network output and target across the training set. Our measure of agreement between the two resulting binary time-series was the φ coefficient. φ is the Pearson correlation coefficient between two binary variables [22]. We also report sensitivity (s_n = TP / [TP+ FN], where TP and FP are the proportions of true and false positive samples respectively, and TN and FN are the true and false negative sample proportions), specificity (s_p = TN / [TN + FN]) and positive predictive value (ppv = TP / [TP + FP]). Overall performance was assessed with leave-one-out cross-validation, in which the data from one subject was set aside and used to test a network trained using the remaining data. This was done once for each of our 8 subjects. The entire 8-fold cross-validation was then repeated 5 times with different initial random weights. Results reported here are means across those 5 cross-validation repetitions. We assessed performance for each subject by calculating mean φ across both their sessions.

III. Results

A. Lapse Rating Model Performance

The lapse rating model output was strongly correlated with the mean human rating (phi correlation, φ = 0.89; Spearman correlation, ρ = 0.81). The strength of this relationship was similar to that between individual human ratings and the mean human rating (mean φ = 0.93; mean ρ = 0.84). Fig. 3 shows a histogram comparing the frequencies of each rating level in the test set. The mean rating and model rating were both quantized to the nearest rating level. The overall distribution is similar, with the model generating more Definitely Not Lapse categorizations (rating = 1) than the human raters. There was good agreement (80%) in the Definite Lapse category. Poorer agreement in Unsure and Possible Lapse categories (levels 3 & 4) are probably largely miscategorizations into neighboring levels (judging by the good overall correlation). Fig. 4 shows an example of the behavior of the model on a typical epoch from the test set.

B. Lapse Detector Performance

Average agreement between the binary network output and the target (rating model output > 3) was moderate, though performance varied substantially between subjects

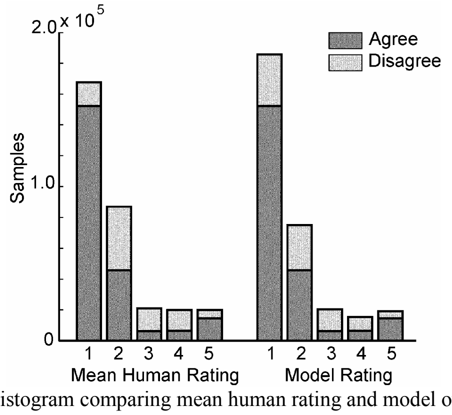

Fig 3. Histogram comparing mean human rating and model output on the test set. The mean rating and model output were both quantized to the nearest rating level.

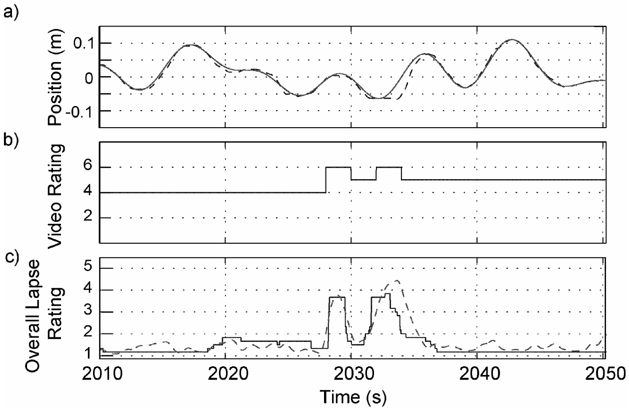

Fig 4. Typical performance of rating model based on a combination of a) tracking behavior, comprising target (solid line) and response (dashed line) and b) expert video rating. c) The rating model output (dashed line) is close to the mean opinion of the human raters (solid line). Time is shown relative to the start of the session.

(mean φ ± SE = 0.36 ± 0.06, range 0.19 to 0.63). The system was moderately sensitive (mean s_n = 0.48 ± 0.09, range 0.14 to 0.83), highly specific (mean s_p = 0.93 ± 0.02, range 0.80 to 0.99) but exhibited poor positive predictive value (mean ppv = 0.39 ± 0.06, range 0.07 to 0.71). Low ppv is tolerable in a lapse detection system, as false alarms have low cost and are preferable to missed lapses.

It was notable that performance on some individual sessions was very good. The best session in the test set gave φ = 0.76, s_n = 0.89, s_p = 0.89 and ppv = 0.81. Fig. 5 shows a typical output of the network, prior to application of the detection threshold.

IV. Discussion

Our results show LSTM can be used to detect lapses, though the detector is not yet sufficiently reliable for general use. Some individual sessions showed reasonably good performance for reasons that are not clear. It could be that these sessions included a greater proportion of deep lapses, making them easier to detect. As with other similar systems, our principal difficulty remains inter-subject variation in the EEG characteristics of behavioral microsleep. In future we intend to identify EEG features facilitating good lapse identification with the aim of improving the reliability of the detector. Work is also underway investigating more robust alternatives to the power spectrum for parameterizing the EEG, with the intention of improving inter-subject reliability.

References

[1] P. Philip, F. Vervialle, P. Le Breton, J. Taillard, and J. A. Horne, "Fatigue, alcohol, and serious road crashes in France: factorial study of national data," *Brit. Med. J.*, vol. 322, pp. 829-30, 2001.

[2] "Grounding of the US Tankship Exxon Valdez on Bligh Reef, Prince William Sound Near Valdez, AK, March 24, 1989," National Transportation Safety Board, Washington, DC 1990.

[3] D. de Waard and K. A. Brookhuis, "Assessing driver status: a demonstration experiment on the road," *Accid. Anal. Prev.*, vol. 23, pp. 297-307, 1991.

[4] S. Porcu, A. Bellatreccia, M. Ferrara, and M. Casagrande, "Sleepiness, alertness and performance during a laboratory simulation of an acute shift of the wake-sleep cycle," *Ergonomics*, vol. 41, pp. 1192-202, 1998.

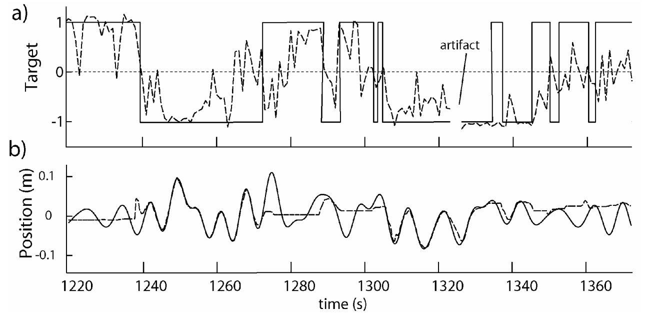

Fig 5. Example of lapse detector performance. a) Detector output (dashed line) and target (solid line) range between 1, indicating a definite lapse, and -1, indicating definitely not a lapse. b) Corresponding tracking behavior with target (solid line) and response (dashed line).

[5] S. K. Lal and A. Craig, "Driver fatigue: electroencephalography and psychological assessment," *Psychophysiology*, vol. 39, pp. 313-21, 2002.

[6] Y. Harrison and J. A. Horne, "Occurrence of "microsleeps' during daytime sleep onset in normal subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 98, pp. 411-6, 1996.

[7] S. Makeig and T. P. Jung, "Tonic, phasic, and transient EEG correlates of auditory awareness in drowsiness," *Brain. Res. Cogn. Brain. Res.*, vol. 4, pp. 15-25, 1996.

[8] S. Makeig and M. Inlow, "Lapses in alertness: coherence of fluctuations in performance and EEG spectrum," *Electroencephalogr. Clin. Neurophysiol.*, vol. 86, pp. 23-35, 1993.

[9] L. Torsvall and T. Akerstedt, "Sleepiness on the job: continuously measured EEG changes in train drivers," *Electroencephalogr. Clin. Neurophysiol.*, vol. 66, pp. 502-11, 1987.

[10] G. Kecklund and T. Akerstedt, "Sleepiness in long distance truck driving: an ambulatory EEG study of night driving," *Ergonomics*, vol. 36, pp. 1007-17, 1993.

[11] R. S. Huang, L. L. Tsai, and C. J. Kuo, "Selection of valid and reliable EEG features for predicting auditory and visual alertness levels," *Proc. Natl. Sci. Counc. Repub. China B*, vol. 25, pp. 17-25, 2001.

[12] J. C. Miller, "Batch processing of 10,000 h of truck driver EEG data," *Biol. Psychol.*, vol. 40, pp. 209-22, 1995.

[13] R. D. Ogilvie, "The process of falling asleep," *Sleep. Med. Rev.*, vol. 5, pp. 247-270, 2001.

[14] T. P. Jung, S. Makeig, M. Stensmo, and T. J. Sejnowski, "Estimating alertness from the EEG power spectrum," *IEEE Trans. Biomed. Eng.*, vol. 44, pp. 60-9, 1997.

[15] A. Vuckovic, V. Radivojevic, A. C. Chen, and D. Popovic, "Automatic recognition of alertness and drowsiness from EEG by an artificial neural network," *Med Eng Phys*, vol. 24, pp. 349-60, 2002.

[16] S. K. Lal, A. Craig, P. Boord, L. Kirkup, and H. Nguyen, "Development of an algorithm for an EEG-based driver fatigue countermeasure," *J. Safety Res.*, vol. 34, pp. 321-8, 2003.

[17] M. T. R. Peiris, R. D. Jones, G. J. Carroll, and P. J. Bones, "Investigation of lapses of consciousness using a tracking task: Preliminary results," in *Proc. EMBC 2004*, San Francisco, 2004, pp. 4721-4724.

[18] D. M. Wolpert, R. C. Miall, J. L. Winter, and J. F. Stein, "Evidence for an error deadzone in compensatory tracking," *J. Mot. Behav.*, vol. 24, pp. 299-308, 1992.

[19] M. H. Hayes, *Statistical digital signal processing and modeling*. New York: John Wiley & Sons, 1996.

[20] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Comput.*, vol. 12, pp. 2451-71, 2000.

[21] R. C. O'Reilly, C. K. Dawson, and J. L. McClelland, "PDP++ Neural Network Simulator," 3.1 ed, 2003.

[22] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: CRC Press, 1997.