Fatigue Recognition Based on Audiovisual Content

Dmitrii Malov
Laboratory of Autonomous Robotics Systems
Saint-Petersburg Institute for Isnformatics and Automation
Saint-Petersburg, Russia
malovdmitrij@gmail.com

Olga Shumskaya
Laboratory of Autonomous Robotics Systems
Saint-Petersburg Institute for Isnformatics and Automation
Saint-Petersburg, Russia
shumskaya.oo@gmail.com

***Abstract*—We analyzed modern approaches to fatigue recognition. Most of the methods are based on classical statistical analysis. Some of the methods are based on machine learning models, in particular artificial neural networks. We have tested independent models for every modality: images and audio and concluded there should be united model for both source of information.**

Keywords—fatigue recognition, neural networks, emotion recognition, audiovisual content

I. Introduction

Fast-paced life, multitasking, irregular working hours, sedentary tech-related work and other factors constantly influence human condition and overall health. Efficiency and safety of human activity directly depend on individual vigilance level, awareness and mental performance. Human factor at work often causes great losses, and frequency of such unfortunate cases accrues together with employee fatigue. Current research, aimed to reveal human fatigue and sleepiness are primarily concerned with car driving process. But equally important is to discover human fatigue in the course of professional duties. To cut employer costs on sick leaves and unscheduled prolonged work missing, it is important to reveal fatigue ahead-of-time, i.e., reliably detect sub-critical fatigue levels. Fatigue and sleepiness detection methods as a whole are divided into obtrusive and unobtrusive ones.

Obtrusive (aggressive) methods assume attachment of electrodes to human body, using special helmets to measure brain waves or special lenses to detect sight focus and eye movement. A downside of such methods consists in usage of cumbersome device, limiting human motions, what can complicate otherwise easy driving, distract and disturb the subject.

Unobtrusive methods assume recording driver on camera and measuring such parameters as steer wheel motion and lateral position of the motor. The advantage of these methods is, that no additional devices are required. The defining factors of human fatigue are nictation length, closed eyes percentage, head rotation activity, nodding

II. Review

Authors of [1] note, that respiration and heart rate are the most distinctive fatigue characteristics. Authors of [2] used high-resolution human face photographs to determine human fatigue. They note, that it is enough calculating linear regression for 8 factors, that show high correlation with fatigue: heavy eyelids, reddish eyes, dark under-eye circles, pale skin, downturned mouth, puffy eyes, glazed eyes, wrinkles and lines around eyes, skin defects or eczema, as well lip tension. This approach assumes that for every face under consideration we define face landmarks using Face++ API, then, based on them, we extract areas of interest, further, using VGG16, calculate features from the obtained areas. Authors note, that the left and the right eye, as well the left and the right eyelid are studied as separate areas of interest.

In paper [3] authors try to reveal early fatigue in drivers. As area of interest the authors chose eye region and set six landmarks per each eye to detect fatigue. Authors tracked nictation as fatigue factor. Within this research a number of existing nictation frequency methods were used; it was noted, that common downside in all of them is the low sensitivity to frequent nictation, where several nictation acts are perceived as a whole, what also was taken into account in this work. Authors also considered in their approach, how well the subject accommodates to changing lighting conditions, moderate head rotations and changing facial expressions. A 10-minute video is used as source data; having processed this video, authors obtain an event vector, where each event contains data on nictation: duration, amplitude, speed and rate of eye opening. The model used here is a HM-LSTM network. Accuracy of fatigue detection in this method was 65.2%.

The paper [4] conceptually describes a prototype of security system, used in car driving. It consists of three self-contained attention modules:

1. Module of safe car cruise in the lane. The main requirement for this module to work: the road should be marked. The module works based on road marks and using a neural network it defines an area, suitable for car cruising.
2. Driver condition module, ensuring driver face recording to detect head posture, nictation, closed eyes, as well recording from driver cab to assess driver’s posture and to know, how his head, body and arms are positioned.
3. Vehicle interaction module, where authors, using a specific device connect to the vehicle systems and obtain data from control units and external sensors, as well video cameras.

Authors of [5] use Dlib library to locate 68 landmarks on the driver's face based on video frames, which are normalized and fed into the neural network. In the experiment several specific settings were considered: driver in 87.1% has been achieved in experiments with driver without glasses, the minimum accuracy – 75.1% – has been achieved in sunglasses, average detection accuracy was 80.9%.

The reported study was funded by RFBR according to the research project № 18-29-22061.

978-1-7281-4911-0/19/$31.00 ©2019 IEEE

An example of obtrusive approach is shown in [6]. Driver fatigue level is determined using a device with 16 electrodes, put on driver's head and taking the encephalogram. Authors consider and calculate the following main parameters:

1. Wakefulness or ratio between alpha and beta signals. Beta-signal corresponds to wakeful state or active mind, alpha-signal corresponds to movements, while the person remains restful; hence, this ratio corresponds to the body activity character.
2. Behavior valence or character, ratio of alpha- and beta-signals, coming from electrodes, attached to the head at prefrontal lobe.
3. Ratio of beta-signals to alpha-signals, coming from electrodes, attached to frontal area of the head.

Each parameter is assessed and fed into fuzzy logic system.

The system, described in [7], includes 4 sequential subsystems:

1. Recording of driver face on camera, installed on the car's dashboard; in this case video is divided into samples.
2. Image preprocessing (noise elimination and contrast optimization), image conversion to grayscale, face detection using Viola-Jones classifier, delineating areas of interest (eyes, mouth).
3. Authors take into account such factors as fast nictation or heavy eyelids, yawning. Then the obtained factors are fed to support vector machine and get fatigue level at output.
4. The fatigue level is classified as "no fatigue" (output "-1"), "low fatigue", "high fatigue" (output "1"). By low fatigue the system generates a sound signal, stimulating the driver; by high fatigue the system triggers water sprinkling and stops the car.

In paper [8] authors study human fatigue due physical activity. Clinical data, obtained in real-time mode, are used as respective parameters. The diagnostics includes timeframe analysis, frequency response analysis, detrended oscillation analysis, approximate and standard entropy. Experiment results show fatigue detection accuracy of 98.65%.

Authors of [9] propose to determine fatigue level using speech recognition and calculating MFC coefficients. This work is based on research, concerning coefficient value changes depending on speaker's fatigue level, coefficients, assigned to certain phonemes, as well coefficients in Sleep Onset Latency test. Authors noted accuracy improvement of 20% relative to prior research.

Also in [10] there is proposed to analyze the voice, determining fatigue level in such manner. Authors use MatLab and PRAAT instruments to extract voice, bearing on such parameters as intensity, transmission frequency, formants, throughput, sound frequency, speech duration, expected value, mean square deviation, standard deviation, energy, power, speech quality, MFC coefficients, mean autocorrection, speech gap duration, etc. Data, calculated this way, are further classified in this paper using: neural networks, SVM, K-means classifier, naïve Bayes classifier and logistic regression. The best outcomes were obtained using SVM: 98.9%.

III. Experiments

We used as database an open video dataset DROZY [11]. This database includes 10-minute long video-recordings, featuring 11 female and 3 male speakers (3 recordings each) for fatigue estimation based on Karolinska Sleepiness Scale – KSS for each recording. It also contains data on manually tagged recordings (screenshots, coordinates of 68 facial landmarks), data on automatic coordinate detection of 68 respective landmarks on each video sample, intervals between some stimuli and human reaction on them, as well polysomnography signals. Each recording corresponds to a psychomotor vigilance test – PVT [12] in the following conditions:

1. Three hours after wake-up.
2. 21 hours after wake-up and 15 hours without caffeine.
3. 29 hours after wake-up and 15 hours without caffeine.

A. *Siamese neural networks*

In the first approach the coordinates of 68 landmarks from each video sample were taken as source dataset. Data preparation consists in composition of 10 different coordinate pairs, taken from arbitrarily chosen samples from every two videos. If the person on both videos is tired, this pair is denoted with "1", else with "0". Because the fatigue examples in the database is tagged according to KSS from 1 to 9, it was decided to treat all the scores less than "6" as "non-tired", whereas "6" and more as "tired". The pairs, prepared such way, are fed into the Siamese neural network; each pair element is an input data for a self-contained neural network, whose output values are calculated using Constructive Loss function [13]:

$$(1-Y)12(DW)2+(Y)12\{max(0,m-DW)\}2, \qquad (1)$$

where DW – Euclidean distance between output values of the output values of Siamese networks.

The Siamese architecture is required not to classify the images, but to find differences between them. Hence the classification loss function (e.g. cross-entropy) is a suboptimal choice here. The Constructive Loss function estimates, how well the network discerns any given image pair.

12600 coordinate sets were composed, where 10080 comprised the training sample and 2520 the testing sample. Experiments showed fatigue detection accuracy of 64.23%.

Apart from a relatively low method accuracy it is important to note, how much time it is required for preprocessing of data pairs. Should any new coordinate set be in place and we need to determine, to which class this data belongs, then it would be necessary to associate the pairs of this with other pairs, already present in the database. With database scaling processing time and RAM volume, needed to do this, will increase respectively.

B. *Model VGG19*

The second approach is based on a pre-trained convolutional neural network VGG-19 [14]. The model was trained on a dataset, composed of more than 1 million images, taken from the ImageNet database. The network contains 19 layers and it can classify images into 1000 categories. The network learned a wide range of feature-based representations for an extensive image set.

In the second scenario screenshot from video playbacks are used as a dataset. Face regions were extracted from these images using dlib library, then the images were normalized. 709 face images were prepared: 534 for model fitting and 175 for testing purposes. Some experiments with multi-classification were performed, where 8 fatigue stages were (the database contained no records with fatigue score “1”), as well with binary classification – scores 1-5 correspond to the class “non-tired”, 6-9 – “tired”. The accuracy of this model was 0.13 on validation dataset. We can conclude, that there is no useful information in face landmarks for this kind of task.

In modern studies, it is noted that mel-frequency cepstral coefficients – (MFCC) and mel-spectral coefficients provide data on the frequency and compressed amount of information that limits the number of calculated coefficients for data processing.

The calculation of the MFCC consists of several stages:

1. Fourier expansion:

$$FFT[k]=\sum x[n]\cdot e-2\cdot\pi\cdot i\cdot k\cdot n\cdot NN-1n=0,\ 0\le k<N. \qquad (2)$$

2. “Smoothing” the values at the borders of frames by multiplying the Fourier coefficients by the Hamming window function:

$$H[k]=0.54-0{,}46\cdot\cos2\cdot\pi\cdot kN-1. \qquad (3)$$

3. Calculation of mel-filters.

Mel – a psychophysical unit of pitch based on subjective perception by average people, that is, a value that determines the significance of sound of a certain frequency.

The frequency conversion to mel is carried out according to the formula:

$$M=1127\cdot\log(1+F700). \qquad (4)$$

Depending on the predetermined number of desired Mel-coefficients and the range of frequencies of interest, the reference points of the filters are determined. The frequency range is converted to the mel calculus, divided into the number of intervals, respectively, the number of chalk coefficients, after which the values of the reference points (interval boundaries) h are converted back to frequency values. The resulting scale is superimposed on the spectrum of the frame:

$$f(i)=[(frame.size+1)\cdot\square(i)sample.rate]. \qquad (5)$$

Now you can define mel – filters:

$$Hm(k)=\{0,\ k<f(m-1)k-f(m-1)f(m)-f(m-1),\ f(m-1)\leqslant k\leqslant f(m)f(m+1)-kf(m+1)-f(m),\ f(m)\leqslant k\leqslant f(m)0,\ k>f(m+1)\}. \qquad (6)$$

4. The use of filters consists in pairwise multiplication of filter values with spectrum energies. It is believed that the logarithm of the resulting Mel coefficients reduces the sensitivity of the coefficients to noise. At this stage, we get the mel-spectrogram.

$$S[m]=\log(\sum|FFT[k]|2\cdot Hm[k]N-1k=0),\ 0\le m<M. \qquad (7)$$

5. Discrete cosine transform (DCT) is used to obtain “cepstral” coefficients, that is, to compress the results, increase the significance of the first coefficients and reduce the significance of the latter.

$$C[l]=\Sigma S[m]\cdot\cos(\pi\cdot l\cdot(m+0.5)M),\ 0\le l<M,\ Mm=0. \qquad (8)$$

IV. Conclusion

The first approach offers the search for the minimum Euclidean distance between pairs of sets of MFC coefficients. The second approach is based on Siamese neural networks, at the input of which pairs of sets of MFC coefficients are fed.

A chalk spectrogram is an option of representing an audio signal in a graph format (the X axis is time, the Y axis is frequency), it is also an informative representation of the signal and can be analyzed for classification. The third approach is based on image analysis of Mel spectrograms of signals with further training of the VGG-19 model and its further application for data classification.

As informative features in the matter of determining a person’s fatigue by voice, the pronunciation speed, reaction rate, and pronunciation accuracy are often singled out. However, it is worth noting that often people can and in a quite peppy state communicate quite slowly, as well as have speech defects or certain features.

Based on the results obtained during the experimental analysis of images, as well as the obvious shortcomings of the audio signal as the only modality for determining human fatigue, in the future work it is planned to combine two modalities: a stream of images and sound.

Acknowledgment

The reported study was funded by RFBR according to the research project № 18-29-22061.

References

[1] A. Chellappa and R. Ezhilarasie, “Fatigue Detection Techniques: A Review,” International Journal of Pure and Applied Mathematics, vol. 117, no. 16, pp. 503–510, 2017.

[2] X. Peng, J. Luo, C. Glenn, L.-K. Chi, and J. Zhan, “Sleep-deprived Fatigue Pattern Analysis using Large-Scale Selfies from Social Media,” In 2017 IEEE International Conference on Big Data (Big Data), pp. 2076–2084, 2017.

[3] R. Ghoddoosian, M. Galib, and V Athitsos, “A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–10, 2019.

[4] H. A. Khojasteh, A. A. Alipour, E. Ansari, and P. Razzaghi, “An Intelligent Safety System for Human-Centered Semi-Autonomous Vehicles,” arXivID: 1812.03953, 2019. URL: https://arxiv.org/pdf/1812.03953.pdf

[5] R. Jabbar, Kh. Al-Khalifa., M. Kharbeche, W. Alhajyaseen, M. Jafari, and Sh. Jiang, “Real-time Driver Drowsiness Detection for Android Application Using Deep Neural Networks Techniques,” Procedia computer science, vol. 130, pp. 400–407, 2018.

[6] M. B. Dkhil, A. Wali, and A. M. Alimi, “Drowsy Driver Detection be EEG Analysis Using Fast Fourier Transform,” In 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 313–318, 2015.

[7] R. Gupta, K. Aman, N. Shiva, and Y. Singh, “An Improved Fatigue Detection System Based on Behavioral Characteristics of Driver,” In 2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE), pp. 227–230, 2017.

[8] M.-Y. Wu, Ch.-H. Chen, and Ch.-Ch. Lo, “An Exercise Fatifue Detection Model Based on Machine Learning Methods,” In 2006 IEEE International Symposium on Signal Processing and Information Technology, pp. 567–571, 2006.

[9] J. Picone and J. L. Berg, “Detecting Fatigue From Voice Using Speech Recognition,” In 2006 IEEE International Symposium on Signal Processing and Information Technology, pp. 567–571, 2006.

[10] M.D. Singh, “Fatigue Detection Using Voice Analysis,” 2015 (Doctoral dissertation).

[11] Q. Massoz, Th. Langohr, C. Francois, and J. G. Verly, “The ULg Multimodality Drowsiness Database (called DROZY) and Examples of Use,” Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV 2016), pp. 1–7, 2016.

[12] M. Basner and D. F. Dinges, “Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. Sleep,” vol. 34, no.5, pp. 581–591, 2011.

[13] H. Gupta, “One Shot Learning with Siamese Networks in PyTorch,” 2017. URL: https://hackernoon.com/one-shot-learning-with-siamese-networks-in-pytorch-8ddaab10340e

[14] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXivID: 1409.1556, 2015. URL: https://arxiv.org/pdf/1409.1556.pdf