MENTAL WORKLOAD ESTIMATION ON FACIAL VIDEO USING LSTM NETWORK

Boris Shishov[1]

[1]Automated Control Systems Department, Gubkin Russian State University of Oil and Gas

(National Research University), Leninsky Prospekt, 65, 119991, Moscow, Russia, E-mail: shishov.b@gubkin.ru

ABSTRACT

To measure the performance of the human-machine system, the human performance must be taken into account. However, it remains difficult to build a well-generalized model of the human condition. Thus we need to focus on modeling of the separate factors. Mental workload estimation is the crucial task in system design and its performance analysis. There are several existing approaches for estimating mental workload, but most of them are not yet appropriate enough for practical use. In this study, the method of estimating mental workload is proposed. The idea is to estimate workload from the video sequence from a common camera (i.e. webcam) using recurrent neural networks (RNN) trained individually. First, the workload is estimated under special experimental conditions using task-based approaches while facial, and eye movement features are extracted from the video captured during the experiment. Then, extracted workload is modeled based on the extracted training data with RNN with long short-term memory (LSTM). As a result, low-error LSTM models are trained.

• INTRODUCTION

Nowadays, most production systems can be defined as an overall combination of human and machine in which both interact close with each other. Obviously, overall system performance and reliability highly depend on the both of the components. However, modeling a machine's performance can be relatively straightforward, but estimating human performance in considerably more difficult. It appears hard to develop comprehensive and computational models of human behavior and performance due to the overwhelming complexity of the subject. Despite a large amount of research is done in the field, it still seems impossible to develop a fully generalized and robust model of human performance. Thus, many researchers focus on the specific human factors and develop models for specific tasks suitable for practical system development.

One of the most important indicators of human cognition is the mental workload. The workload can be characterized as a mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, coupled with the capability of the operator to respond to those demands. The main reason for measuring mental workload is to quantify the mental cost of performing tasks to predict operator and system performance. It especially matters in environments where increased task demand may lead to unacceptable system performance.

There are three main categories or the approaches for measuring mental workload: subjective rating scales (e.g. questionnaires), performance measures (based on task completion performance) and psychophysiological measures. Different categories are sensitive to different aspects of the mental workload leading to a situation where different measures are assessing slightly differing things, mostly due to the lack of the strict definition of the workload. Subjective rating scale based models, i.e. NASA-TLX (task load index) [1] often produces a special index of mental demand after certain activities is already done by the operator. Thus, this kind of models is not suitable for a long-term real-time monitoring. There is a vast variety of task-performance based models which are often presented as generative or predictive models, but they are highly dependent on the detailed description of operator's behavior regarding tasks and goals. There are notable examples of these models: GOMS family models [3], Model Human Processor (MHP) [4], ACT-R [5], and even queuing network models like QN-MHP [2]. However, all of these models requires detailed tasks breakdown to simple operators, which is possible only when the task is already well known. In complex, partially observed or unknown environments, such models might be unsuitable. The other category of approaches relies on the psychophysiological measures, which allow extracting certain measures in dynamic without the need to describe the task. However, these methods often require a strong and robust translation of certain signals to modeled variable but often suffers from signal's insensibility to the task workload. Most commonly used the psychophysiological channel to measure workload is electroencephalography (EEG) [6]. However, there are many studies that rely on other channels like heart rate (and its other measures), respiratory rate, blink rate, eye movement, facial temperature and even eye pupil dilation. However, these methods often require additional hardware like EEG measuring devices, or high-speed infrared cameras for eye gaze tracking and thus may be unsuitable for practical use.

In our approach, we combine both the task-based modeling approach and psychophysiological approach relying on the recent advancements in machine learning. We propose a model to estimate the mental workload from the facial video, since head pose, facial expression [10] and eye gaze [9] gives us much information about the operator's both physical and mental state. There are robust models for modeling facial expressions like Facial Action Coding System (FACS) [15], which is producing numerical estimation regarding Action Units (AUs). Strong capabilities of the modern computer vision techniques [7-10] allow us to use the consumer level camera (e.g. webcam) as the only additional hardware to capture facial features. In the current study, we tried to create an estimator of the mental workload from the FACS Action Units and eye gaze information [16]. This estimator can be considered as the time-series prediction model given a set of time-series (facial features). To achieve that we decided to use recurrent neural networks (RNN) with long short-term memory (LSTM) [11] since these models have proven to be effective [12] in time-series analysis. To estimate the workload we first need to train our LSTM model by providing it with some target variable (mental workload). To capture "ground truth" mental workload the special experiment is conducted in which mental workload is modeled by detailed task analysis.

• THE PROPOSED METHOD

The basic idea behind our method is to train an LSTM [1] RNN model with the facial features (including FACS Action Units and eye gaze information) extracted from the video and the target mental workload. The training data (i.e. facial features and workload) is captured during the special experiment. To estimate mental workload during experiment task analysis with adjusted MHP [4] model is performed, since we know exactly which tasks participant performs during the experiment. Once the model is trained, it can be used for estimation and prediction of the mental workload just from the video from a common webcam. To achieve this, we need to perform multiple steps. Each of the steps is explained in the following subsections.

• Experiment design for estimating workload

To obtain the training data, the special experiment which models mental workload must be conducted. After research of the existing task-based models [2-5], we decided to design the experiment to be very simple to reduce overall modeling system complexity. Participants are asked to complete primitive tasks on the computer – to click on the circles on the screen. These circles appear in the random places on the screen and disappear after fixed amount of time (1300 milliseconds). Circles are of the color and initial shape; also, circles are increasing in radius during lifetime (from 30 to 40px). The circle appears with the delay, which is changing during the experiment. Starting with delay equal 1200ms and then for the 3 minutes this delay decreases to the 300ms making tasks appear faster over first 3 minutes. Starting from 3rd minute till minute 4 the delay increases back to an initial value (1200ms) making the task appear slower, which leads to relaxation of the participant after very intense tasks. The exact formula used to calculate the delay is as follows:

$$delay(t) = \begin{cases} 1200 - (1200 - 300) * \dfrac{t}{180000}, t \leq 180000 \\ 300 + (1200 - 300) * \dfrac{t - 180000}{60000}, t > 180000 \end{cases}$$

Where t – experiment time in milliseconds. During the experiment session (4 minutes overall) participant's task performance is captured. It consist of reaction time (time from circle appearance until being clicked, in milliseconds), distances from the mouse clicks (required for further motor time estimation), missed click events, non-clicked circle events. In addition, a video sequence of the participant's face in captured.

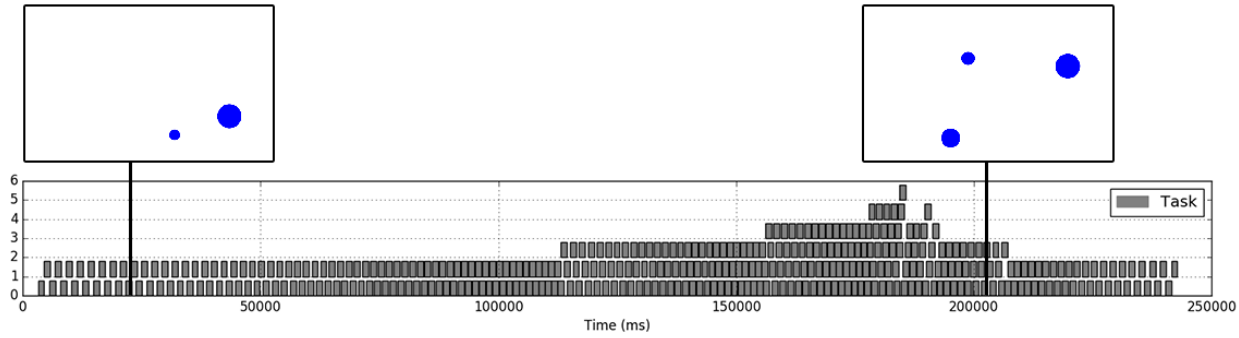

Figure 1: Graph of appearing tasks during experiment and screenshots at different moments. Each gray box represents a circle lifetime (1300ms) on the screen. In peak, there are up to five circles on screen at one moment.

Despite tasks being primitive, clicking circles could still be appropriate for modeling workload since that according to MHP [4] even these tasks still require perceptual (circle appears on the screen), cognitive (decision to move the cursor) and motor (moving the cursor) efforts. Thus, we need to break down each task into these phases. The tasks are almost the same by design, which leads to almost the same amount of time required for each phase for different tasks. However, the participant used to move eyes and the cursor for a random distance each task (since circles appear at random position on the screen) which affects the time required for perceptual and motor response. It is hard to measure eye movement and perception time with enough precision, so we have to assume (according to MHP [4]) that perception of each circle is nearly equal to the perceptual processor cycle time (100ms). However, the time required for motor response is larger than perceptional, and as we know exact mouse positions over the time, we can apply Fitts's Law derivative for two-dimensional tasks [13] to estimate motor response time. After the motor time is estimated for each task, we then remove time overlapping trying to keep motor time as close as possible to the estimate because in this experiment human cannot perform several motor responses (moving mouse) at once. Each task does not require memory efforts, as the circle clicking is mostly just a sensory-motor operation, which leads to simplified MHP model. As a result, we got that reaction time (RT) is a sum of perceptual time (assuming 100ms), cognitive time and motor time, which is modeled with the Fitts's Law. The example of such breakdown is shown in figure 2.

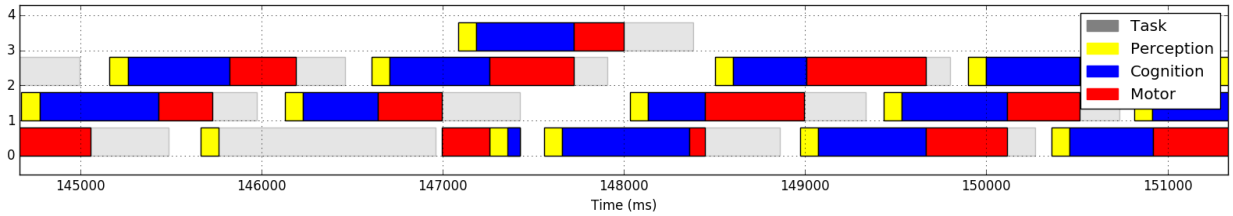

Figure 2: Tasks reaction time breakdown. Tasks consist of perception (yellow), cognition (blue) and motor response (red). Motor responses are not overlapping. There is a missed click event at the 147s, which requires some time for verification after. There is a missed task at 145.8s, which only produces perception effort.

Once this breakdown is performed, we could estimate the cognitive workload dynamics through the cognitive effort. To do that we stack vertically cognitive bars from the graph ignoring the overlapping we captured (because overlapping caused by a psychological refractory period (PRP) [14] phenomena). As a result, we get the graph where filled parts mean that at that time participant showed some cognitive effort (Figure 3, top). This process can be formalized as follows:

$$effort(t) = \begin{cases} 1, if\ \exists\ T_i \in Tasks: t \in [T_i^{start} + P_i, T_i^{start} + RT_i - M_i] \\ 0, otherwise \end{cases}$$

where t – time (in milliseconds), T_i – one of the tasks from the whole experiment, task start time is the time when circle appears on the screen, RT_i – reaction time of that task, P_i – time required for perception, M_i- time required for motor reaction.

As we calculated the time regions of the workload, we could now estimate the mental workload itself. We might think of the time as of the resource then total time of cognitive effort during some period divided by the period will give us resource load index which is mental workload. The idea is to slide a fixed time window across the effort, calculating mental workload as follows:

$$workload(t) = \left(\int_{t-window}^{t} effort(t)\, dt \right) / window$$

Where $window$ – interval length in milliseconds. For this experiment, we used a window equal 10000 milliseconds; see results at Figure 3 (bottom).

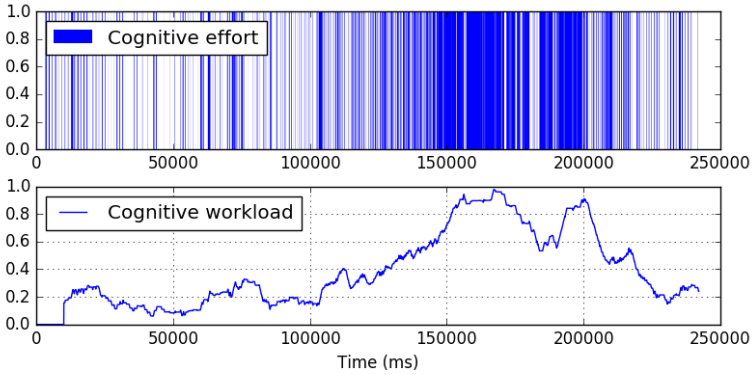

Figure 3: Processed task execution results obtained from the experiment. Top: cognitive effort visualization. Each filled time region means that participant showed some cognitive effort at this time. Bottom: cognitive (mental) workload, produced by aggregating cognitive effort.

The computed cognitive workload curve then will be used for training the model. Also, it is possible to perform similar steps to calculate perception workload and motor workload curve.

- **Extracting facial features from the video sequence**

To obtain training data, besides workload curve, we also need to extract facial features from the video captured during the experiment. The facial expression itself contains much information about the human state. There is a lot of the research is done in this area. The main approach to describe facial expression is to apply the Facial Action Coding System (FACS) [15]. It provides a detailed explanation of facial expression regarding Action Units (AU), that could be numerically evaluated, which gives us the ability to construct models on top of it. Recent advancements in computer vision [10] produced robust models to extract Action Units from video sequences. These models rely on facial landmark estimation models [8]. Another useful feature to model workload could be eye movement. It is proven [16] that eye movement could be a reliable feature for estimating mental workload. In our model, we applied existing model [9] for extracting such information.

Our goal in this process is to extract facial and eye-movement features from captured video sequence. Since this problem is already well studied, we developed an extraction software on top of OpenFace toolkit [7]. The features that we extract from the video are: head pose information (position and rotation in 3D) eye gaze positions (two 3D points, one for each eye) and intensity 17 Action Units resulting 29 features in total per frame. The average camera framerate is 25 frames per second and

experiment lasts for 4 minutes, which results in 6000 frames. This means that for each participant as the features for training we got 29 time-series 6000 samples long.

- **LSTM RNN model**

Once all the training data is obtained i.e. mental workload curve and extracted features from the video, we need to train the model to estimate workload from facial features. Both mental workload and facial features are a time series, which means our goal is to develop the time-series prediction model. One of the most powerful models for such tasks are recurrent neural networks (RNN) with long short-term memory (LSTM) [11, 12], we will use them for estimating workload. RNN networks differ from regular neural networks by a different architecture: RNN networks contains memory element, while regular ones do not. This makes RNN better for sequence analysis since they can use information about previous input samples, while regular neural networks work just with the current sample.

Network architecture (Figure 4) used for estimating workload contains an LSTM layer with 32 output units and 0.2 dropout to prevent overfitting followed by a dense layer with only one output (mental workload). Input data at each training step has a shape of 29 features (scalars) while output shape is a single scalar.

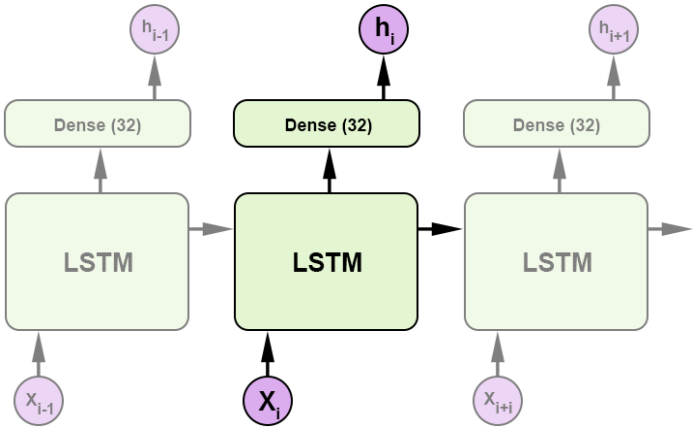

Figure 4: Network architecture. It consists of two layers: LSTM and Dense layer (32 input units resulting one output). X – is the input features, h – is the output workload estimate.

In our recent research [17] we have shown that facial expressions differ from person to person while performing cognitive tasks. Thus, we will focus on training models for each person separately considering its differences. The model was implemented in Python using Keras deep learning framework [18]. Training was done in 5 epochs with batch size equal 64 samples. It takes about 2 minutes to train the model on one participant's data. Model evaluation can be performed with more than 60 samples per second.

- **EXPERIMENT RESULTS**

The experiment was conducted with 20 participants (15 male, 5 female with an average age of 20). Each participant was asked to sit in front of the computer and to complete the whole sequence of tasks (clicking circles for 4 minutes) in front of the camera. The only restriction for participants was not to talk. The test itself was implemented in HTML and JavaScript.

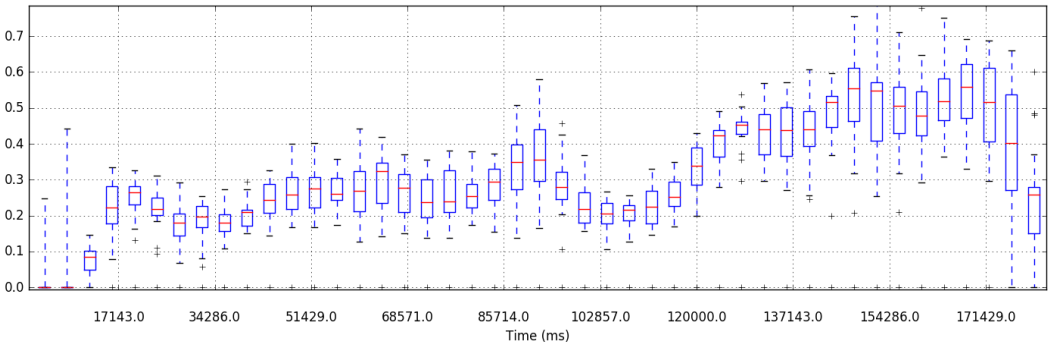

Figure 5: Boxplot of all participants workload during the experiment. All of the participants produced similar workload dynamics as the test was the same for all of them. However, each participant had its own performance curve.

As the results, we obtained a video sequence and test events for each participant, which then processed to facial and eye movement features and mental workload estimated curve. Each participant's result displayed a strong correlation between all results of the estimated mental workload (Figure 5) since the test is designed to force cognitive effort. Some of the participants

displayed notable changes in facial expression during the experiment (Figure 6). Which confirms the existence of the connection between facial expression and mental workload.

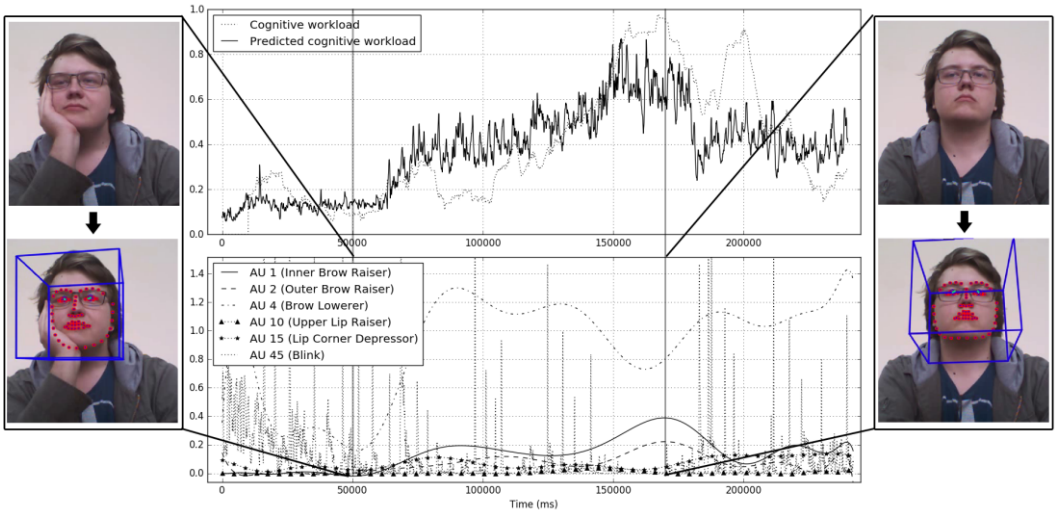

Figure 6: Obtained results for one of the participants. The chart on the top depicts workload modeled with the task-based approach (dashed line) and the predicted workload by LSTM model. The chart on the bottom shows several extracted features (only 6 of 29 total). Images on the sides are the frames from the video at the given moment.

Once these results are obtained, the described model was trained for each participant separately. LSTM RNN models displayed great performance modeling mental workload sequence from video features resulting low error values – average mean squared error (MSE) for all participants is 0.02 (with std 0.008). As we can see in Figure 6 – the trained LSTM model reacts well to considerable changes in facial expression.

• CONCLUSIONS

As the result of this study, the approach is proposed for estimating mental workload for primitive but intense tasks. Which is originally based on the MHP model, but modified to be more accurate with simple but concurrent tasks along with the formulation of estimation method of mental workload itself. To test this approach, the special experiment was designed and conducted with 20 participants. For each participant individually a proposed LSTM model was trained. The result trained models showed good performance predicting mental workload from the extracted features.

The overall approach includes workload estimation from known tasks-based environments, facial and eye movement features extraction from the video sequence and applying LSTM RNN models to estimate workload from extracted features. In theory, this approach could be easily integrated for practical use since no special hardware is required – the only required hardware is the camera which could be a common webcam.

However, this approach still requires some more research: tests in different environments i.e. different tests which imply stronger cognitive activity like arithmetical computation; comparison with different task-based models of mental workload (i.e. QNMHP) and a long-term testing or monitoring.

Appendix

All of the source code used for this study available at https://github.com/bshishov/MentalWorkload and https://github.com/bshishov/Emotions.

References

[1] Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*. P.A.M. Hancock, N. Amsterdam, North-Holland: 139-183.

[2] Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queueing Network-Model Human Processor (QN-MHP): A computational architecture for multitask performance in human-machine systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *13*(1), 37-70. Approach for Modeling Cognitive Performance. *Human Factors and Ergonomics Society Annual Meeting Proceedings* 45(24):1733-1737

[3] B. John and D. Kieras (1996) The GOMS Family of User Interface Analysis Techniques: Comparison and Contrast. *ACM Transactions on Computer-Human Interaction*, Vol. 3, No. 4, December 1996, Pages 320 –351.

[4] LL, L. G. (1986). An engineering model of human performance. *Month*, *1*, 35.

[5] Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

[6] Roy RN, Charbonnier S, Campagne A, Bonnet S. (2016) Efficient mental workload estimation using task-independent EEG features. *Journal of Neural Engineering*, Vol.13, No.2

[7] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency (2016) OpenFace: an open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision, 2016*

[8] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency (2013) Constrained Local Neural Fields for robust facial landmark detection in the wild. *IEEE Int. Conference on Computer Vision Workshops, 300 Faces in-the-Wild Challenge*, 2013

[9] Tadas Baltrušaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling (2015) Rendering of Eyes for Eye-Shape Registration and Gaze Estimation Erroll Wood, *IEEE International*. *Conference on Computer Vision (ICCV)*, 2015
[10] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson (2015) Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. *Facial Expression Recognition and Analysis Challenge, IEEE International Conference on Automatic Face and Gesture Recognition*, 2015
[11] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
[12] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*.
[13] MacKenzie, I. S., & Buxton, W. (1992, June). Extending Fitts' law to two-dimensional tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 219-226). ACM.
[14] Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological review*, *104*(4), 749.
[15] Ekman, P., & Friesen, W. V. (1977). Facial action coding system.
[16] May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta psychologica*, *75*(1), 75-89.
[17] Shishov, B. A., Kolesnikova, A. S., & Poulin, M. A. (2017). Development of a psycho-emotional state model of a manager of an automated control dispatching system (ACDS). *Automation, telemechanization and communication in oil industry*, (1), 15-23.
[18] Chollet, F. (2015). Keras.