Workload estimator using EEG and eye-tracking

Ivan Tashev
Microsoft Research
Redmond, WA, USA
ORCID 0000-0002-2263-2047

Christine Beauchene
MIT Lincoln Laboratory
Lexington, MA, USA
ORCID 0000-0003-0787-4520

R. Michael Winters
Microsoft Research
Redmond, WA, USA
ORCID 0000-0002-8874-9184

Yu Te Wang
Microsoft Research
Redmond, WA, USA
ORCID 0000-0001-5576-5236

David Johnston
Microsoft Research
Redmond, WA, USA
davidjo@microsoft.com

Nathaniel Bridges
Air Force Research Laboratory
Dayton, OH, USA
ORCID 0000-0001-9899-6165

Justin R. Estepp
Air Force Research Laboratory
Dayton, OH, USA
ORCID 0000-0001-8049-9582

***Abstract*—In this paper we present a workload estimator based on biological signals - electroencephalographic and eye-tracking. The workload estimator is person- and session- independent, designed to work in a virtual reality flight simulator environment and is a part of our adaptive training system. The novel component is using objective evaluation of the workload, based on the flight logs, as labels for training the regression neural network. As evaluation parameter is selected the correlation with the objective labels. The paper contains the results from using several feature sets and estimators, where the best estimator achieves correlation with objective labels of 0.84.**

***Index Terms*—workload estimation, electroencephalography, eye-tracking, adaptive training system**

I. Introduction

Mental workload is a subjective measure of the cognitive demands imposed by a task on an individual. It can affect performance, safety, and well-being in various domains, such as aviation, education, and health care. Therefore, it is important to develop reliable and valid methods to assess mental workload in real-time and in natural settings. One promising approach is to use physiological signals, such as electroencephalographic (EEG) and eye-tracking, that can reflect the cognitive state of the user [1].

EEG and eye-tracking can provide complementary information about mental workload, as they capture different aspects of cognitive processing [2]. EEG signals reflect the electrical activity of the brain, which can be analyzed in terms of frequency bands, such as alpha, beta, theta, and gamma. These bands are associated with different cognitive functions, such as attention, memory, and problem-solving [3]. Additionally, eye-tracking features such as fixation duration, saccade amplitude, and blink rate, can indicate relevant changes in attention, engagement, and workload based on task demands [2], [3].

However, estimating mental workload using EEG and eye-tracking also poses several challenges, such as individual variability, session-to-session variability, and task specificity. In many cases, the labels used for training the estimators are noisy – either based on self-evaluation or based on the task performed, which can cause different workload based of the person's experience. In many cases, the process of mental workload measurement requires individual calibration and a strictly controlled environment. Also, the output of the estimator is very granular, frequently limited to low, medium, and high mental workload.

In this paper, we propose a mental workload estimator for the flight simulator environment that produces mental workload estimations in the form of a score from 0 to 100, where 0 is very high workload and 100 is very low workload. The estimators are trained using an objective performance measurement, derived from the flight logs of the simulator. Using these objective low noise labels allow training of neural-network-based estimators that outperform the traditional linear regression and Support Vector Machine (SVM) estimators.

II. Experimental setup

A. Adaptive training system

The adaptive training system (ATS) [4] is a human-in-the-loop system aiming to accelerate the training process for pilots using virtual reality (VR) flight simulators, as shown in Fig. 1. The training process is assumed to go through short, indivisible, scenario runs, called trials, that end with a score – a number describing how successful the run was. Based on the model of the training process [5] and the past scores the system determines the parameters describing the trainee (initial absolute skill level and the learning speed) and recommends the scenario difficulty for the next run, optimal in sense of maximal increase of the absolute skill level. Then a scenario with the closest difficulty is selected from a larger library of scenarios. In this model, each scenario is characterized by its difficulty and the maximum achievable score. The scenario difficulty is increased by adding wind, thermals, wind gusts, and decreasing the visibility from clear to fog resulting in fully instrumental flight. The total number of scenarios with different difficulties was eleven. The simulation analysis, using statistical models of the subjects, showed up to a 25% reduction of the training time. If the training time is fixed, then ATS leads to maximizing the final average absolute skill level.

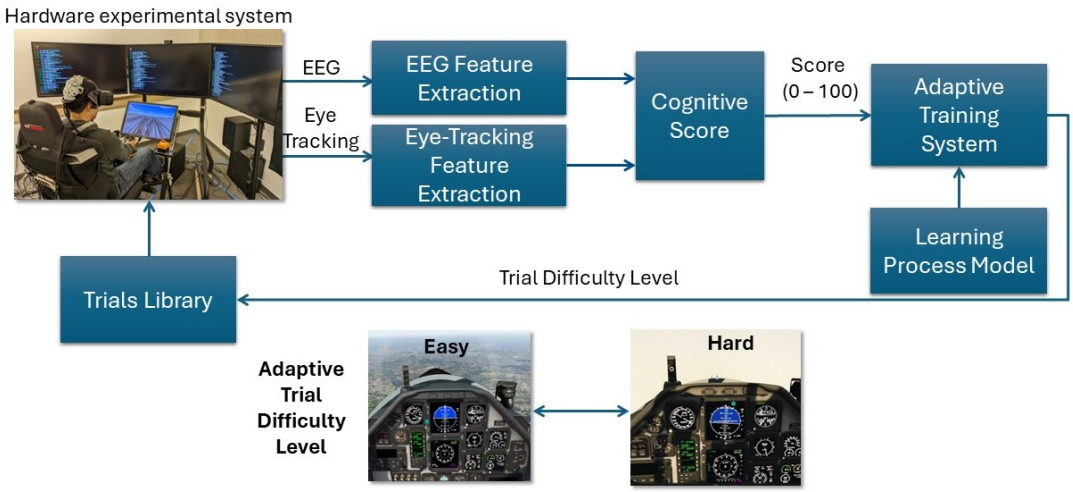

Fig. 1. Block diagram of the Adaptive Training System (ATS).

B. Scoring as critical component of ATS

Proper assessment of the workload is critical for an efficient adaptive training system. In this context, the workload is represented by the score. Note that the score aggregates both the scenario difficulty and the subject's skill level. Presumably when they match the subject will receive a score somewhere in the middle of the range 0-100. If the scenario is more difficult than the skills of the subjects then ATS produces a lower score, representing higher workload. In the other case, when the scenario is easier than the skill level of the subject, ATS produces a higher score, representing lower workload.

In this specific data collection, the general scenario is a straight-line flight in two cases: a) straight and level flight; b) glideslope flight. In the first case, the subject is expected to maintain at a constant altitude, speed, and course. In the second case, the subject is expected to maintain at a constant course and speed while flying the plane towards the beginning of the runway. In both cases, we can retrieve from the logs of the flight simulator the position of the plane at any moment of the flight. Each trial has a duration of 2-3 minutes. Then we can estimate the root mean squared error (RMSE) between the desired and the actual trajectory of the plane and use it as a score after normalization between 0 and 100. This way of scoring has its own problems and has been improved [6]. Still, this way of scoring is applicable only for simple tasks when we know the "right" trajectory of the plane.

Using biological signals to estimate the workload is applicable for a much broader range of tasks and is the subject of this paper.

C. Experimental hardware system

The experimental hardware system consists of flight simulator Prepar3D running in virtual reality (VR) mode. The subject is sitting on a 6 degree of freedom (DoF) motion platform, wearing Varjo VR3 VR glasses. The VR glasses are equipped with custom-made 32 dry EEG electrodes, plus a ground and reference. These electrodes are connected to the BrainVision LiveAmp pre-amplifier with a 24 bits ADC and 500 Hz sampling rate. Eye-tracking is integrated into the Varjo VR3 headset. In addition, from the subject one channel electrocardiographic, galvanic skin resistance, and breathing signals are collected.

III. Data collection and feature extraction

A. Data collection

Each subject participates in the data collection process over five consecutive days within one work week. In the first day, the subject is instructed to sign the IRB (reviewed and approved by the ethical committees at Microsoft Research and the Air Force Research Lab (AFRL)) and answer pre-experiment questionnaires, followed by ten trials. In the next four days, the subject runs twenty trials per day. The maximum number of trials per subject is 90, some of these might be disqualified due to various reasons – signal quality being the main of them. There were several days when the subject did not show up as well.

The data collection was conducted in three waves. The first was following the protocol above with 6 participants using only the two easiest scenarios for straight-and-level flight and glideslope. The second wave involved 10 subjects without flying experience running all 11 scenarios in random order. All 10 subjects were asked every day before the sessions to sit and relax in the chair with open eyes for two minutes doing nothing. Their physiological signals (EEG and eye-tracking) were recorded and treated as a regular scenario run with an assigned score of 100 (no mental workload). During the preliminary data processing, we noticed that the subjects have predominantly zero scores on the most difficult scenarios. To obtain better information about these scenarios, the third wave of data collection was with 9 subjects with a valid pilot license. They were invited only for one day and executed 20

TABLE I
COMBINING ELECTRODES IN GROUPS.

Group	Electrodes
F_L	AF7, F3, F5
C_L	C5, C3, C1
P_L	CP5, CP3, CP1
F_R	AF8, F4, F6
C_R	C2, C4, C6
P_R	CP2, CP4, CP6
PO	P5, Poz, P6

runs each. The number of scenarios recorded and used for further processing is 1223 from total of 25 different subjects.

The data collection process, questionnaires, and restrictions on participants were reviewed and approved by the ethical committees at Microsoft Research and the Air Force Research Lab (AFRL).

B. Data pre-processing and feature extraction

All the flight logs were processed and the scores for each run computed. Then we ran the modeling of the training process, according to [5] and determined the parameters of the scenarios (difficulty level and maximum achievable score) and of the trainees (initial absolute skill level and learning rate). Using the final subjects and scenarios parameters was conducted a simulation for each subjects and simulated scores estimated. These scores are much smoother, as they are based on all scores from all subject from all scenarios. We used these scores as labels for training the estimators.

The raw EEG data was processed using MNE [7] to re-reference and bandpass filter between 1 - 55 Hz to remove drift and powerline noise was applied. Then, each channel was epoched into one second long frames, and multiple statistical measures (mean, range, kurtosis) of the signal were computed. The epochs with severe interference due to eye-blinking, high noise, or motion artifacts were removed. Then for the clean signals, the power of the delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-60 Hz) bands were estimated at 5 second increments. Then, we reduced the number of the extracted power features by averaging across channels. The electrodes used to form the frontal (F), central (C) and parietal (P) groups on the left and right side and the reference group are shown in Table I. For each session this leads to a vector with variable length with 35 features in each element (5 bands x 7 channel groups). The length of the vector depends on the duration of the session and the number of frames rejects and is typically 80-120 elements.

The eye-tracking data was processed in a similar manner. After converting the Varjo data to PyTrack [8] format, the data was processed to extract various oculomotor features over 30 second epochs (e.g. fixation duration, fixation dispersion, gaze entropy, and blink rate), forming another variable length vector with 39 features for each epoch. Both EEG and eye-tracking data were cropped to align with the start and end of the trial. However, due to modality specific pre-processing steps, the length of the eye-tracking feature vector is similar to the length of the EEG features vector, but not necessarily identical.

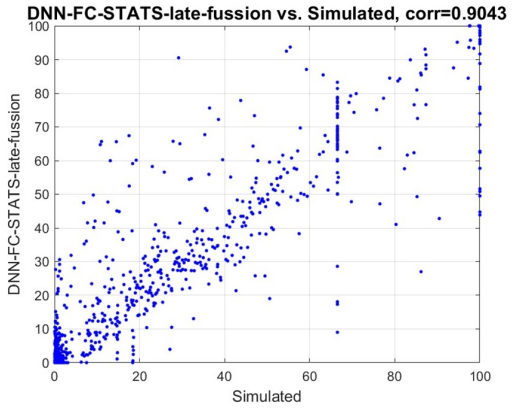

Fig. 2. Simulated vs. estimated results.

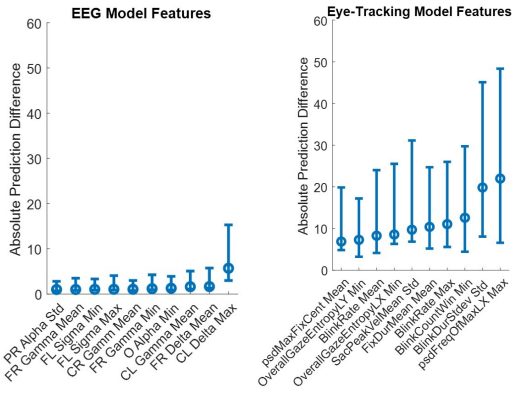

Fig. 3. Ablation study results.

Most of the estimators are expecting a fixed number of features. We have one label per session and two vectors with variable and different length features from EEG and eye tracker. To overcome this problem, we do average across the timeline and compute the mean, the deviation, the maximal and the minimal values of each feature. This means that for EEG we have 7 channels x 5 bands x 4 stats = 140 features for each session. For the eye tracker data, we have 39 features x 4 stats = 156 features for each session. The total number of features is 296 for each session and one label – the estimated score.

IV. ESTIMATORS AND TRAINING PROCESS

We treated the scoring problem as a regression machine learning problem – for each session we have a set of features and a label. As an evaluation criterion, we selected the correlation coefficient between the simulated scores (i.e. the labels) and estimated by the regression engine scores. As we target person- and session- independent estimator, we remove one subject to be used for testing, one subject to be used for validation, and use the rest of the subjects for training. For more reliable verification and testing we used only subjects with full number

TABLE II
RESULTS, CORRELATION WITH THE SYNTHETIC SCORES

Features		Validation set					Test set				
		LIN	SVM	ELM	DNN	LSTM	LIN	SVM	ELM	DNN	LSTM
mean of original feature set	EEG	**0.1471**	0.1204	0.1469	0.1375		**0.1471**	0.1204	0.1381	0.1332	
	EYE	0.4592	0.1615	0.2509	**0.4816**		0.4592	0.1615	0.2437	**0.4693**	
	Fusion	0.7207	0.7240	0.7359	**0.7560**		0.7207	0.7240	0.7331	**0.7403**	
	Early fusion	0.4392	0.1596	0.1488	**0.4814**		0.4391	0.1496	0.1414	**0.4511**	
mean, max min, std of the original feature set	EEG	0.1143	0.1250	0.1182	**0.1290**		0.1143	0.1250	0.1058	**0.1186**	
	EYE	0.4243	0.3262	0.2909	**0.5449**		0.4233	0.3263	0.2790	**0.5417**	
	Fusion	0.8093	0.8073	0.8394	**0.8402**		0.8093	0.8073	0.8397	**0.8376**	
	Early fusion	0.2816	0.3417	0.3399	**0.4688**		0.3786	0.3253	0.2743	**0.5087**	
Sequence	EEG					0.3173					0.2986
	EYE					0.5613					0.5499
	Fusion	0.8390	0.8464	0.8384	**0.8489**		0.8390	0.8464	0.8384	**0.8371**	

TABLE III
ABLATION STUDY, PER GROUP OF EEG FEATURES: BAND AND ELECTRODES

Algorithm	Baseline	Bands					Electrodes groups						
		Delta	Theta	Alpha	Beta	Gamma	F_L	C_L	P_L	F_R	C_R	P_R	PO
LIN	0.0544	0.0285	0.0321	0.0216	0.0607	0.0644	0.0628	0.0288	0.0494	0.0448	0.0999	0.1142	0.0733
SVM	0.1078	0.148	0.0754	0.0995	0.1204	0.1155	0.1023	0.1111	0.1174	0.1024	0.1083	0.1214	0.1245
ELM	0.0724	0.0603	0.061	0.0659	0.0983	0.0558	0.0787	0.0348	0.0963	0.0749	0.1234	0.1391	0.0978
DNN-FC	0.0920	0.1094	0.0848	0.0708	0.1112	0.1087	0.0567	0.0726	0.072	0.0556	0.0832	0.1264	0.0414
Average	0.0816	*0.0865*	**0.0633**	**0.0644**	*0.0976*	0.0861	**0.0751**	**0.0618**	0.0838	**0.0694**	*0.1037*	*0.1253*	0.0842

TABLE IV
ABLATION STUDY, PER GROUP OF EEG FEATURES: STATISTICS

Algorithm	Baseline	Mean	Std	Min	Max
LIN	0.0544	0.0747	0.0559	0.0737	0.0304
SVM	0.1078	0.1274	0.1149	0.1128	0.0896
ELM	0.1078	0.1000	0.0896	0.0576	0.0212
DNN-FC	0.0920	0.0929	0.0953	0.0709	0.0583
Average	0.0816	*0.0987*	*0.0889*	**0.0787**	**0.0499**

of scenarios and scores – 90. In our data set there are five. This means that we have 20 combinations of subjects used for testing and validation. All results provided further in this paper, are average correlation coefficients from all 20 training and evaluation combinations.

We have two groups of features – from EEG and from the eye-tracking. We can have two approaches for estimation of the score: a) early fusion, when all features are combined into one feature vector and one estimator is trained; b) late fusion, when we train one estimator for each group of features and one for fusing the outputs of these estimators for the final score estimation.

In both approaches we have experimented with the following estimators:

- Linear regression, which is the straightforward estimation using least squares method [9].
- Support Vector Machine (SVM) in regression mode [10].
- Deep Neural Network (DNN) with a given number of layers and nodes in each layer [11].
- Extreme Learning Machine (ELM) in regression mode, which is a shallow and wide neural network with one hidden layer and analytic solution for the training [12].

In addition to these estimators, we used a neural network that preserves a state, like long-short term memory neural network (LSTM). Then each element of the variable length feature vector is turned into a feature, and we take the output after the final element. Because of the different length of the feature vectors from EEG and eye-tracking for the sessions only late fusion is a feasible option here.

The training of the DNN and LSTM was limited to 150 iterations with forced stopping if the results on the validation data set did not improve for five epochs. The stochastic gradient descent algorithm was used for training with an initial learning rate of 0.001. The hyper-parameters of each estimator (number of layers, number of nodes, etc.) were optimized for each regression strategy using the average correlation on the validation datasets. This resulted in 64 nodes in the ELM hidden layer with sigmoid activation, three layers of 32 nodes for the regression DNN, 128 nodes for the LSTM. To ensure repetitiveness of the results and reliable optimization of the hyper-parameters we start each training and evaluation run with a fixed seed of the random number generator.

V. RESULTS

The results from all approaches for the validation and test sets are shown in Table II. The first group of the lines shows the results from using only the means for each feature,

the second group of lines are the results from using mean, deviation, min and max values. The last group of lines show the results from the LSTM network. Notably, in all cases the late fusion approach outperforms by far the early fusion approach. In majority of the cases DNN structure outperforms the linear regression, SVM and ELM. Notable also is that the EEG and eye tracker estimators, based on LSTM, show better individual results, but after the final fusion we have practically the same correlation with the simulated labels as with the DNN-based estimators. Based on these results and the stability and repetitiveness of the results from the various combinations of validation and test sets we would recommend using the late fusion approach with DDN for EEG, eye tracker, and late fusion estimators. The results from all data (train+validation+test) are illustrated in Fig. 2.

VI. Ablation study

We applied an ablation study to investigate the impact of various EEG and eye-tracking features on the recommended above approach for late fusion with DNN for all three estimators. Therefore, we systematically removed each feature to identify the primary drivers of the model (see Fig. 3). For EEG the primary features are in the Delta band, which may be associated with attention. For eye-tracking the primary features are related to the frequency of eye movements and blinking, which could be associated with attention and fatigue. For predicting performance, the eye-tracking measures have a larger effect on the prediction compared to the EEG. This may be because the eye-tracking data is less person- and session-dependent compared to the EEG signals.

Another point of view provided the ablation study per EEG feature groups: bands, electrodes groups (Table III), and feature groups (Table IV). By removing the features from each group, we found that the most useful bands are alpha and theta, while beta and delta reduce the results least, which is consistent with previous finding of cognitive workload [13]. The most useful electrodes are in the frontal and central parts (F_L, F_R), while parietal electrodes on the right are less useful (P_R, C_R, PO). The results from the processed groups show that max and min are most useful, followed by deviation and mean.

VII. Conclusions and future work

In this paper we proposed and evaluated a model for predicting simulated scores based on brain and eyes tracking signals. The model is person- and session- independent and does not require a calibration process. The correlation with the simulated scores is above 0.83, which is sufficient for practical applications of the ATS. The main reason for these results is using objective low noise scores as a labels for training the estimators.

Future work includes further refining of the model, based on further feature set analysis to identify key features across all participants. We intend to explore more sophisticated neural networks, CNN and LSTM as examples. We will add the other bio-signals collected (ECG, breathing, GSR) and study their importance in mental workload estimation. It will be very interesting how this trained model will estimate the mental workload outside of the flight simulator environment.

VIII. Acknowledgements

The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense, or of the United States Air Force. The content or appearance of hyperlinks does not reflect an official Department of Defense, United States Air Force, or Air Force Research Laboratory position or endorsement of the external websites or the information, products or services contained therein.

References

[1] C. Mühl, C. Jeunet, and F. Lotte, "EEG-based workload estimation across affective contexts," *Frontiers of Neuroscience*, vol. 8, 2014.

[2] A. Aygun, B. Lyu, T. Nguyen, Z. Haga, S. Aeron, and M. Scheutz, "Cognitive Workload Assessment via Eye Gaze and EEG in an Interactive Multi-Modal Driving Task," in *Proc. of ICMI'22*, Bengaluru, India, Nov. 2022.

[3] S. Shadpour, A. Shafqat, S. Toy, Z. Jing, K. Attwood, Z. Moussavi, and S. Shafiei, "Developing cognitive workload and performance evaluation models using functional brain network analysis," *npj Aging*, vol. 9, no. 1, p. 22, 10 2023.

[4] I. Tashev, R. M. Winters, Y.-T. Wang, D. Johnston, and N. Bridges, "Adaptive training system," in *Proc. IEEE Research and Applications of Photonics in Defense Conference (RAPID)*, Sep. 2023.

[5] I. Tashev, R. M. Winters, Y.-T. Wang, D. Johnston, A. Reyes, and J. Estepp, "Modeling the training process," in *Proc. IEEE Research and Applications of Photonics in Defense Conference (RAPID)*, Sep. 2022.

[6] I. Tashev, R. M. Winters, Y.-T. Wang, D. Johnston, J. Estepp, and N. Bridges, "Towards a better scoring," in *Proc. IEEE Research and Applications of Photonics in Defense Conference (RAPID)*, Sep. 2023.

[7] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, no. 267, pp. 1–13, 2013.

[8] U. Ghose, A. Srinivasan, W. Boyce, and et al., "Pytrack: An end-to-end analysis toolkit for eye tracking," *Behav. Res.*, vol. 52, pp. 2588–2603, 2020.

[9] A. S. Goldberger, "Classical linear regression," *Econometric Theory*, p. 158, 1964.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] G.-B. Huang, Q.-Y. Zhu, and S. C.-K, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[13] K. C. Ewing, S. H. Fairclough, and K. Gilleade, "Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop," *Frontiers in human neuroscience*, vol. 10, p. 223, 2016.