Sensing and Assessing Cognitive Workload Across Multiple Tasks

Matthias D. Ziegler[1(✉)], Amanda Kraft[2,3], Michael Krein[2], Li-Chuan Lo[4], Bradley Hatfield[4], William Casebeer[1], and Bartlett Russell[1]

[1] Lockheed Martin Advanced Technology Lab, Arlington, VA, USA
{matthias.d.ziegler, william.d.casebeer, bartlett.a.russell}@lmco.com
[2] Lockheed Martin Advanced Technology Lab, Cherry Hill, NJ, USA
{amanda.e.kraft, michael.krein}@lmco.com
[3] Drexel University, Philadelphia, USA
[4] Department of Kinesiology, University of Maryland, College Park, USA
{llo, Bhatfiel}@umd.edu

Abstract. Workload assessment models are an important tool to develop an understanding of an individual's limitations. Finding times of excess workload can help prevent an individual from continuing work that may result in human performance issues, such as an increase in errors or reaction time. Currently workload assessments are created on a task by task basis, varying drastically depending on sensors and task goals. Developing independent models for specific tasks is time consuming and not practical when being applied to real-world situations. In this experiment we collected physiological signals including electroencephalogram (EEG), Heart Rate and Heart Rate Variability (HR/HRV) and Eye-Tracking. Subjects were asked to perform two independent tasks performed at two distinct levels of difficulty, an easy level and a difficult level. We then developed and compared performance of multiple models using deep and shallow learning techniques to determine the best methods to increase generalization of the models across tasks.

Keywords: Workload · Cognitive · EEG · HRV · Flight simulator · Neural networks · Predictive models · Deep belief network · Linear SVM

1 Introduction

Over the last decade computational models have gained an increasing presence as techniques to understand human behavior and in linking behavior to physiological measures [1–3]. Studies utilizing computational models have successfully shown links between measures of workload with performance that were not previously apparent due to the large amount of data that current sensors are able to collect [4, 5]. While many studies implement such models they can vary significantly between studies due to the diversity of tasks being tested, number/type of sensors and analysis techniques. This leads to highly specialized models that do not transfer between tasks and individuals.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2016, Part I, LNAI 9743, pp. 440–450, 2016.
DOI: 10.1007/978-3-319-39955-3_41

A model that is so specialize that any change to the task or individual being modeled requires complete system retraining is impractical in applications outside of controlled experiments. To bring computational models outside of the lab for practical use in real world environments it is important to examine how physiological data can be reliably processed and analyzed in a manner that is beneficial for understanding workload levels and performance across both individuals and tasks of interest.

Recently, studies have shown that cognitive workload levels can be measured using an increasing number of available sensing techniques. Electroencephalogram (EEG) has been one of the most common tools for measuring workload, identifying increased neural activity corresponding to workload levels [6–8]. Eye tracking is also common, as evidence of pupil size and blink rate have been linked to workload levels [9, 10]. Electrocardiogram (ECG) offers another means to assess workload levels via heart rate variability [11]. By combining sensors some studies have been able to show workload levels consistent across multiple physiological sensors [5, 12] and to increase the classification accuracy of any one of these systems alone by accounting for a greater number of physiological systems that respond to changes in workload. In this study we use this combined sensing approach to measure performance in multiple tasks to determine how well general levels of workload are linked to task performance across individuals. While other tools, such as fNIRS, fMRI and biomarkers in particular [13, 14], have also shown to be important measures of workload we do not address these tools in this study.

Understanding the ability of computational models to predict performance from physiological signals is an important tool that could have large number applications in cognitively demanding environments. While many studies have looked at linking workload measures with performance and have even shown success predicting future performance [13], there has been no comprehensive study evaluating the possibilities and limitations of what a combination of physiological measures can predict. Here we start that process by looking at creating a single subject agnostic generalizable model to predict performance based on EEG, eye tracking and ECG. We compare accuracy of a model that is trained based on the performance results over two independent tasks and then tested on a separate hold-out population the model was not trained on. We compare these results with a model that is trained on a single trial from all individuals and tested on the same population over multiple additional trials as well as a model trained and tested on an equal, random distribution of all available data. We posit that subject's physiological signals are unique enough that unless their performance is represented in the training algorithm there will be a decrease in model accuracy of performance prediction. However the exact tradeoff between drop-off in accuracy and individualization necessary for adequate performance is unknown. This study plays an important role in understanding difficulties that occur when trying to use physiological data as a measure of performance without individualized training.

Computational models can vary in complexity of programing, amount of data needed for training, time needed to run the model and number of parameters that need to be adjusted (i.e. layer size, learning rates, etc.) It is important to examine how the accuracy of different modeling approaches affects predictions in an effort to understand the tradeoff between accuracy and time needed to develop/execute the model. In this study we analyzed the results of two types of model, but focus primarily on one a deep

belief network using neural networks. We chose a neural networking approach to model human performance, based upon the ability of neural networks to robustly classify nonlinear data. Additionally we did some initial testing comparing the neural network with a linear support vector machine (SVM) model. In each model we did comprehensive cross-validation by randomly distributing the subject pool into training and testing sets and running the model 25 times to ensure accurate reporting the performance of each model. Simply running a model with a single training/testing set may cause skewed results as the training or testing data chosen may not be representative of the overall data. As performance variation between cross-validation runs is an indicator of dataset variability and an estimate for overall method reliability with respect to the data, we will present the models results we tested over the 25 model cross-validation runs and indicate that cross-validation should be standard procedure when testing models.

The result of our study shows that a generalized model with no tailoring performs very poorly when applied to a new individual and that some level of model adaptation or personalization is necessary for any level model prediction to be valid. These findings provide an indication that a comprehensive study is necessary to understand the trade offs between generalized model performances versus the costs of adapting models to individuals.

2 Methods

2.1 Experiment Setup

Participants. A mix of thirty-five right-handed undergraduate and graduate students from University of Maryland were trained and tested on two computer based video games to measure performance over a period of four non-consecutive days. Subjects were trained on the system during day 1 and tested days 2, 3 and 4. Four trials (each from different participants) were disregarded due to errors in recording.

Task Design. We designed two tasks to titrate workload: a simple Snake Game (see Fig. 1) and Prepar3D Flight simulator (Fig. 2), each of which contained multiple levels of difficulty. Subjects received one 45 min training period to become familiar with the tasks and returned for 3 days following the training to perform the tasks. The order of the tasks and difficulty levels were randomized for each day. Each difficulty level lasted for 5 min in both tasks.

2.2 Task Description

Nokia Snake Game. A video game was developed using Presentation programming language to mimic the Nokia Snake game preloaded on Nokia cellular phones. The game, shown in Fig. 1, consists of a "snake" that moves at a constant pace, but the subject controls the snake's direction with keyboard commands, including up, down, left or right. The goals of the game are to avoid hitting any walls or the snake itself (in

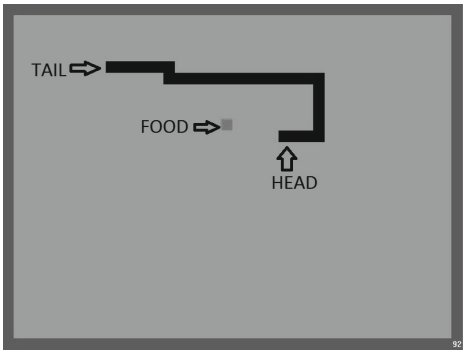

Fig. 1. Example of the "snake game", the subjects control the head to eat the food while avoiding its own tail and the surrounding walls (grey boarder).

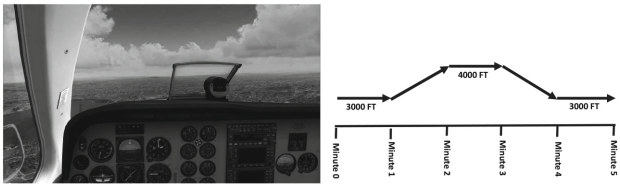

Fig. 2. Example seen from Prepar3D cockpit (left) task instructions (right).

which case the snake "dies" and the level restarts), and to collect as much "food" as possible. When subjects direct the snake to "eat" the food, the food adds a single additional square to the snake's length There is no limit to the snake's length, but as it grows longer, it becomes more difficult to navigate within the maze without hitting a wall or itself. The subjects completed two different levels of Snake: "easy" in which the snake moves at a slow speed (traveling across the screen only once every 100 ms) and a faster "hard" speed (moving once per 38 ms). The game provides an element of automatic titration to player skill; the length of the snake increases the difficulty of the subject's ability to eat food while avoiding itself and the walls of the game. During testing we found that in the "hard" condition the keyboard input latency was delayed such that two fast-sequential keystrokes did not always register as the subject intended.

Prepar3D Flight Simulator. Lockheed Martin's Prepar3D flight simulator was used as a second performance task. Subjects were given control of an aircraft and the task was broken down into 5-one minute subtasks. During minute one they were asked to maintain level flight at an altitude of 3000 ft. with a heading of 180 degrees at speed of 180 knots. In minute two, they were asked to maintain the same direction, but increase their altitude to 4000 ft while never increasing altitude at rate less than 1000 ft/min. They then maintained 4000 ft for another minute before decreasing again to 3000 and maintained that altitude for the final minute. To create multiple difficulty levels the environment in which the plane was flying was changed. In the "easy" condition there was no wind and no turbulence, however in the harder condition winds of over 30

knots and severe turbulence affected the aircraft's position, causing high levels of difficulty maintaining the desired heading and altitude.

Physiological Tools. During testing days subjects' physiological signals were measured using BrainVision EEG system with electrodes arranged according to the standard 10–20 system [15]. Additional BrainVision electrodes placed on the collarbone recorded electrocardiogram (ECG) for HRV analysis. An SMI eye tracker recorded eye movements and pupillometry.

2.3 Computational Models

Model Development and Performance Estimation. We wanted to understand the impact of traditional signal processing methods on our ability to develop transferable human performance models. We baseline corrected the data using the average Welch's Power Spectral Density (PSD) computed from each individual's resting state "eyes-open" session and removed the data for the first and last 10 s for each trial. We computed the PSD for each task, using Welch's method over 1 s interval of data. For each channel and frequency bin, the corresponding baseline PSD average was subtracted from the task PSD. Each resulting 1-second interval was used as a unit of data for training or testing during model development and validation.

The data was tested in two models a Linear SVM model (shallow model) and a neural network model (deep belief network). The deep belief network structures were Gaussian-Bernoulli Restricted Boltzmann Machine (GB-RBM) classifiers based upon Tanaka's code [16]. We estimated appropriate learning rates, learning step sizes, hidden layer sizes, and drop rates based on successive modeling performance during automated model tuning. We tuned models via sequential grid tuning approach; for a particular layer size and drop rate, a grid of step rates (0.1, 0.2, 0.3….) was evaluated. A step rate is then selected based on highest mean AUC and is used for subsequent modeling. The process is repeated for determining drop rates (0, 0.25, 0.5, 0.75…). For all models developed, the layer size chosen for use in determining the optimal learning rate was 100; the drop rate used for selection of optimal learning rate and hidden layer size was 0.5. All human data were standardized via z-score scaling prior to modeling.

Data Analysis. For each of the preprocessing strategies outlined above, Receiver Operating Characteristic (ROC) curves were analyzed and areas under the ROC curve (AUCs) are reported. An AUC of 0 refers to no prediction ability of the model, 0.5 denotes chance prediction, and 1.0 denotes perfect prediction of the model. Here we report the average and standard deviation of AUCs based upon 25 rounds of cross validation. The predictions are based on a binary decision for the model. The model predicts if the current test data that is provided to the model is above or below the median score for the task.

Cross Validation. In one of our approaches to cross-validation, we randomly assign 75 % of the subjects (all trials) as a training set, and withhold the remaining 25 % of subjects as a validation set. This is key to the approach, and reflects the transferability

of human performance indicators without prior knowledge. It is important to note that since a developed model here has not been trained on any prior data from the validation set individuals, one would expect modest predictive ability. We compared this to other modeling strategies where data were withheld such that some data from each individual was included in both the training and testing set. We tested this by withholding a single trial from each individual randomly for the validation set (and the rest of the trials were part of the training set). In addition, we performed training modeling experiments where a subset of all subjects trials were included in the validation subject's, the balance of data (75 %) remained in the training set.

3 Results

3.1 Performance

Subjects showed distinct performance differences between the easy and hard levels of both the snake game and the flight simulator. The snake score was determined by adding the amount of time the subject stayed "alive" (avoiding running into walls or other parts of the snake) plus the growth of the snake plus the Manhattan distance from the food.

The average performance difference was 0.5572 ± 0.086 for the easy snake game versus 0.4658 ± 0.0745 for the hard snake game, as shown in Fig. 3. The density value in Fig. 3 represents the number of subjects at a particular score. This scoring system was developed after analyzing each individual measure and recognizing that no one measure alone was representative of the overall performance. The combined mean which was used as the binary output for the model was 0.5115. The Prepar3D flight simulator task showed similar distinct differences in the average subject performance; however there was more variance between subjects and trials within the easy and hard conditions. The mean values for the Prepar3D performance was 0.4517 ± 0.2282 for the easy condition and 0.323 ± 0.2188 for the hard condition with a combined mean of 0.3874. These scores represented a normalized analysis of how far from the desired heading, speed, altitude and vertical feet per minute the subject varied. A score of 1 would have indicated 0 drift from the desired measures.

3.2 Model Results

Deep Belief Network. The deep belief networks showed significant differences in performance based on the type of model used (parameters chosen), type of training performed within the model and the specific task. As outlined in Table 1, all of the deep learning models consistently improved based on the type of training used for the model.

Validating the model on subjects that the model was not trained on showed only around chance levels of performance, 0.49 ± 0.4 accuracy in predicting performance for Prepar3D and 0.54 ± 0.04 predictions for the Snake Game when the best parameters

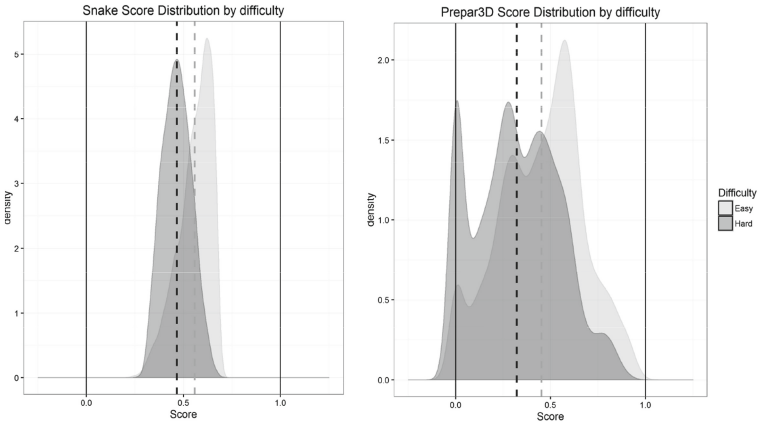

Fig. 3. The snake subject performance (left) on easy (light gray) and hard (dark gray show distinct differences. The Prepar3D task performance (right) had higher variance within a given condition.

Table 1. AUC results for deep belief network and lindear SVM models for varying training/testing datasets (rows) and varying parameters (results of min/max and average shown in the columns

Deep belief network results				
Prepar3D				
	Mean AUC	Min AUC	Max AUC	AUC SD
Split by subject	0.49	0.38	0.58	0.04
Split by trial	0.60	0.57	0.62	0.01
Random split	0.79	0.76	0.81	0.01
Snake game				
	Mean AUC	Min AUC	Max AUC	AUC SD
Split by subject	0.54	0.46	0.59	0.04
Split by trial	0.61	0.58	0.63	0.01
Random split	0.71	0.69	0.73	0.01
Linear SVM				
Prepar3D				
	Mean AUC	Min AUC	Max AUC	AUC SD
Split by subject	0.5234	0.4134	0.6584	0.07
Split by trial	0.4533	0.4027	0.5316	0.05
Snake				
	Mean AUC	Min AUC	Max AUC	AUC SD
Split by subject	0.4981	0.4011	0.5924	0.05
Split by trial	0.4747	0.3966	0.6034	0.07

are chosen (Max AUC). When the training and validation data sets were determined by separation of trials (i.e. inclusion of data from the same three trials for all subjects in the training set and all data from remaining trials in the validation set), the predication ability increased, up to 0.60 and 0.61 when all subjects were included with separate subject trials withheld for validation. The increase in prediction ability peaked with 0.79 and 0.71 when data from all subjects and all trials were represented in both the training and validation of the model.

Linear SVM Model. The linear SVM model was tested as a comparison for only the split by subject training. When training and validation on unique subjects the mean prediction accuracy of the SVM was 0.52 ± 0.07 and 0.4981 ± 0.05 for the Prepar3D and Snake game respectfully. They performed minimally in cross validation of 0.41 and 0.40 for each task and a maximally at 0.65 and 0.59 for each task. There was no significant change in performance when the training was split to including one training session from each subject.

4 Discussion

As scientific researchers push to obtain as much physiological data as possible in order to understand how human performance is holistically defined they will rely on more complex computational models, as traditional data analysis methods will not be sufficient. The results of the experiments in this study show that how the data is processed and modeled can create large variations in overall predictive power. In order to choose the correct computational model, one must take multiple factors into consideration, which we will discuss here.

4.1 Physiological Sensors and Task Design

In this experiment we chose multiple physiological sensors that would record measurements that have been linked to cognitive workload levels. Our original predictions were that as the task became more difficult the workload would increase and performance decrease under the median, while during easy tasks there would low workload and high performance. However the link between workload and performance is not a one-to-one relationship, in fact depending on expertise level it has been shown that two subjects who perform equally may have very different workload measurements [13]. To combat this we trained the subjects for an equal amount of time on multiple tasks in which they had little to no experience in and that drastically differed in necessary levels of workload. For example, during the flight simulator easy task, to maintain constant altitude, heading and speed the subject simply needed to hold the flight stick steady and pull back with minimal force to change altitude. When the flight simulator experiment switched to hard level, independent of expertise, large effort and attention was needed as the winds and turbulence changed the planes course affecting all performance levels agnostic to how well the subject controlled the plane. Similarly the Snake game speed on the easy level was slow enough that avoiding the walls was a simple task and subjects were able to control the snake to gain only enough length that they felt

comfortable with the control. When it was switched to hard the subjects not only had to increase concentration based on the speed of the snake, but as an unattended consequence of our task not recording every keystroke, subjects had to change strategies in real time to compensate for keystrokes not responding, theoretically increasing workload levels. The performance graphs of these tasks showed distinct differences in performance between the easy and hard tasks. While the trend comparing model performances was consistent across tasks, there were significantly greater differences in the flight simulator task over the snake task. This difference can be attributed to multiple aspects of the task, first the flight simulator had distinct quantitative goals (altitude, heading, speed, feet/minute) that every subject aimed for. During the snake task subjects had discretion if their priority would be to stay "alive" or to eat as much food as possible and could obtain the same performance measure. The second limiting factor for the snake game was the altered controls in the hard task which may have caused subjects to change their priorities between the easy task in how they approached the food. In the easy task they could take quick turns, but in the hard task they may have elected for larger turns toward the food to compensate for the delay in response to keystrokes and this may have lowered the score or frustrated some subjects, causing workload independent changes to the physiological signals between task difficulty levels.

4.2 Training Method

Given that we saw very similar performance distributions between subjects on the easy and hard levels in both tasks we predicted we would be able to make a single generalizable model for each task that would work to predict performance on novel subjects. However, when we trained both the deep belief networks and the linear SVM on a 75 % subset of the subjects and tested using the remaining 25 % of the subjects the models performed only at a $\sim$50 % prediction accuracy (validation). Under our 25 model runs, altering which subjects were part of the training and which part of the testing the best results were 65 % for the Linear SVM and 59 % for the Deep Belief Networks. These results showed that while some distribution of the subjects caused the model training to be more representative of the larger population, the models were still poor at predicting performance based off of workload measures. The fact that multiple models showed this performance led us to believe that even with similar performance, there was not a standard workload measure that worked consistently across individuals for either task.

To verify if the model could accurately predict performance from individualized workload measures, we trained the deep belief networks in a number of other ways to determine if we could generate a better performing model. When we trained the models using a subset of each subjects' data the models performance significantly improved for the Deep Belief Network, but we saw no improvement in the Linear SVMs. We did this two ways, first by assigning a single session of each subject to train and tested on the remaining sessions (both models), second we randomly chose data points from each subject across all of their data (deep belief only) allowing each subject and testing session to be represented in both the training and testing model runs. While both of

these model runs significantly improved the performance of the deep belief network, the later showed the greatest improvement accuracy predicting up to ~80 % of the tests in the flight simulator task. The worst cross-validation model performance with this modeling technique was equivalent to or only slightly better than the best general model where there was no training/testing overlap. The model's improved ability to accurately predict performance form physiological workload measures when all subjects are represented in both the training and testing sets illustrates the extreme differences within physiological measures across individuals corresponds to the same behavioral outcomes. Only when a model is personalized for the intended user and possibly to a specific task, will it be reliable and useful as a predictive model.

We posit that the ability of the models to perform best when all sessions are represented in the training session is due to learning that may occur within subject across sessions. Even if performance remains equivalent across sessions, as the subject becomes more familiar with the tasks (and more comfortable wearing all the sensors) their physiological measures may change independent of performance measures. Therefore, not only is it necessary to account for individualized differences when developing a computational model to predict performance or categorize workload, but a model most also account for learning that occurs over time. Even experts in a given field have been shown to change performance and continue learning, albeit at a slower rate, thus models must account for this even in cases when naïve subjects are not being used.

While individualized models show high performance, the time and effort taken to train a new model for every subject is extremely time consuming. To combat this need for individualization, our future work will examine the ability to group subjects based off of current and prior performance and create a set of template models to which a subject can be quickly matched. The template may then be tailored to the individual as the subject improves at the task shortening the overall process. By having not one, but a set of models trained on only subjects that show the same performance trends we posit a high accuracy prediction without a one-to-one relationship between number of models and subjects. This set of models will be the only possibility to create real-time modeling that will be necessary if adjustments to the subjects' performance or tasking are desired in timely fashion.

References

1. Kieras, D.E., Meyer, D.: Computational Modeling of Human Multiple-Task Performance. No. TR-05/ONR-EPIC-16. Michigan University, Department of Electrical Engineering and Computer Science, Ann Arbor (2005)
2. Hugo, J., Gertman, D.I.: The use of computational human performance modeling as task analysis tool. In: Proceedings of the Eighth American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies, NPIC&HMIT 2012, pp. 22–26 (2012)
3. Meng, J., Wu, X., Morozov, V., Vishwanath, V., Kumaran, K., Taylor, V.: SKOPE: a framework for modeling and exploring workload behavior. In: Proceedings of the 11th ACM Conference on Computing Frontiers, p. 6. ACM (2014)

4. Ke, Y., Qi, H., He, F., Liu, S., Zhao, X., Zhou, P., Ming, D.: An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. Front. Hum. Neurosci. 8 (2014)
5. Liu, Y., Ayaz, H., Onaral, B., Shewokis, P.A.: Neural adaptation to a working memory task: a concurrent EEG-fNIRS Study. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2015. LNCS, vol. 9183, pp. 268–280. Springer, Heidelberg (2015)
6. Kamzanova, A.T., Kustubayeva, A.M., Matthews, G.: Use of EEG workload indices for diagnostic monitoring of vigilance decrement. Hum. Factors: J. Hum. Factors Ergon. Soc. **56** (6), 1136–1149 (2014)
7. Walter, C.B.: EEG workload prediction in a closed-loop learning environment. Doctoral dissertation, Universität Tübingen (2015)
8. Brouwer, A.M., Hogervorst, M.A., Van Erp, J.B., Heffelaar, T., Zimmerman, P.H., Oostenveld, R.: Estimating workload using EEG spectral power and ERPs in the n-back task. J. Neural Eng. **9**(4), 045008 (2012)
9. Bodala, I.P., Kukreja, S., Li, J., Thakor, N.V., Al-Nashash, H.: Eye tracking and EEG synchronization to analyze microsaccades during a workload task. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pp. 7994–7997 (2015)
10. Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O.N.M., Atkins, M.S.: Workload assessment of surgeons: correlation between NASA TLX and blinks. Surg. Endosc. **26**(10), 2746–2750 (2012)
11. Ke, Y., Qi, H., He, F., Liu, S., Zhao, X., Zhou, P., Ming, D.: An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. Front. Hum. Neurosci. 8 (2014)
12. Choe, J., Coffman, B.A., Bergstedt, D.T., Ziegler, M.D., Phillips, M.E.: Transcranial direct current stimulation modulates neuronal activity and learning in pilot training. Front. Hum. Neurosci. 10 (2016)
13. Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B.: Optical brain monitoring for operator training and mental workload assessment. Neuroimage **59**(1), 36–47 (2012)
14. Just, M.A., Carpenter, P.A., Miyake, A.: Neuroindices of cognitive workload: neuroimaging, pupillometric and event-related potential studies of brain work. Theor. Issues Ergon. Sci. **4** (1–2), 56–88 (2003)
15. Jasper, H.H.: Report of the committee on methods of clinical examination in electroencephalography: 1957. Electroencephalogr. Clin. Neurophysiol. **10**(2), 370–375 (1958)
16. Tanaka, M., Okutomi, M.: A novel inference of a restricted boltzmann machine. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 1526–1531. IEEE (2014)