

# A Trial of Intellectual Work Performance Estimation by Using Physiological Indices

Shutaro Kunimasa<sup>1</sup>, Kyoich Seo<sup>1</sup>, Hiroshi Shimoda<sup>1</sup>, Hirotake Ishii<sup>1</sup>

<sup>1</sup>Kyoto Univesity, Yoshida-honmachi Sakyo-ku, Kyoto 606-8501, Japan  
{kunimasa,seo,shimoda,hirotake}@ei.energy.kyoto-u.ac.jp

**Abstract.** In order to evaluate the intellectual productivity quantitatively, most of conventional studies have utilized task performance of cognitive tasks. Meanwhile, more and more studies use physiological indices which reflect cognitive load so as to evaluate the intellectual productivity quantitatively. In this study, the method which estimates task performance of intellectual workers by using several physiological indices (pupil diameter and heart rate variability) has been proposed. As the estimation models of task performance, two machine learning models, Support Vector Regression (SVR) and Random Forests (RF), have been employed. As the result of a subject experiment, it was found that coefficient of determination ( $R^2$ ) of SVR was 0.875 and higher than that of RF ( $p < 0.01$ ). The result suggested that pupil diameter and heart rate variability were effective as the explanatory variables and SVR estimation was also effective in task performance estimation.

**Keywords:** Intellectual Productivity · Machine Learning · Physiological Indices · Pupil Diameter · Heart Rate Variability

## 1 Introduction

Recently, mental works such as intellectual works have occupied most of office works in companies and have become more and more valuable in our society. Therefore, economic and social benefits can be larger by improving intellectual productivity such as the efficiency and accuracy of performing intellectual works. In order to achieve this, the quantitative evaluation of intellectual productivity is required, and several studies have been conducted [1]. However, a number of cognitive tasks used in experiments for the evaluation are different from actual office works, because the tasks have been designed in order for experimenters to collect operation logs easily and accurately. In order to evaluate intellectual productivity in actual office, it is desired to measure it when conducting actual office works. It is, however, difficult to collect and evaluate most of their logs. On the other hand, there are various studies which focused on human internal states (e.g. psychological states and cognitive load). Especially, Cognitive Load Theory study (CLT) [2-5] is one of the examples which have a deep relationship with working memory and cognitive load, which have an influence on several intellectual abilities. The goal of CLT is to estimate human internal states by measuring cognitive load in order contribute the design of education and office environment. In this point,

CLT is highly compatible with the evaluation of the intellectual productivity. In order to evaluate cognitive load, various conventional studies [6-9] employed physiological indices (e.g. heart rate and pupil diameter). And they revealed the relationship between these indices and cognitive load. Furthermore, several studies [10,11] performed the estimation of the intellectual work performance by using physiological indices. It is expected that these studies contribute the evaluation of the intellectual productivity of office workers. However, the objective and quantitative evaluation method has not been established yet.

In this study, so as to develop the quantitative evaluation method of the intellectual productivity by physiological response, the estimation method of task performance is considered by using physiological indices measured when performing a cognitive task. As the physiological indices for task performance estimation, heart rate variability and pupil diameter are employed, which are supposed to reflect human cognitive states (e.g. stressed and relaxed). These indices are also expected to be measured by contactless measurement method without giving any load. As an estimation model, two machine learning models, Support Vector Regression (SVR) and Random Forests (RF), have been considered because they are typical methods in the conventional studies which classify several cognitive states by using physiological indices. And the estimation accuracy of these models was compared by using coefficient of determination ( $R^2$ ) as accuracy indicators.

## 2 Physiological Indices

It is known that there is a close relationship between cognitive activities and physiological responses. In this study, contactless-measurable physiological indices were employed in order to estimate task performance because it is undesirable that the equipment of physiological indices measurement gives stress to a worker when measuring these indices. Concretely, pupil diameter [12] and heart rate variability [13] were employed because it is expected that contactless and accurate measuring methods will be developed soon.

### 2.1 Heart Rate Variability

The spectral analysis of heart rate variability has been conducted in various studies because it is supposed to reflect an autonomic nerve response against stress and cognitive load [14,15]. Mulder [16] found that the power of 0.06-0.14Hz of heart rate variability was reduced by difficulty of cognitive task and high load on working memory. Moreover, low frequency (0.04-0.15Hz) of heart rate variability (LF), high frequency (0.15-0.50Hz) (HF) and the ratio of LF and HF (LF/HF) reflect a sympathetic nerve, a parasympathetic nerve and the balance of these nerves, respectively [17]. In this study, these features, LF, HF, and LF/HF, were employed because it is assumed that the change of cognitive load according to task performance gives an impact on the autonomic nerve and it changes these three features (LF, HF, and LF/HF). In order to extract these feature values, the time window was set to 5 minutes (5-minute timeframe) because it should be long enough to contain at least 15-30 cycles of LF to make it accurate to extract 3 feature values of heart rate variability. LFs and HFs were extracted by calculating low

(0.04-0.15 Hz) and high (0.15-0.50 Hz) frequency wave of heart rate in 5-minute timeframe with shifting it every 1 minute. Moreover, LF/HF was also calculated by dividing LF by HF. These feature values were defined as explanatory variables for task performance estimation.

## 2.2 Pupil Diameter

In the field of psychophysiology, it is known that pupil diameter changes under various cognitive states (e.g. stressed and relaxed). There are various conventional studies which proved the validation of pupil diameter as the index of load on working memory [6-9]. On the other hand, there is a study which suggests a relationship between speed of information processing and the size of pupil diameter [18]. Moreover, there is another study which showed that difficulty of sound classification has a relationship with the size of pupil diameter [21]. In this study, pupil diameter, therefore, was employed as an effective feature for task performance estimation. Mean pupil diameter (MPD) was extracted as a feature value of pupil diameter by calculating the average of pupil diameters in 5-minute timeframe with shifting it every 1 minute as well as the heart rate variability.

## 3 Estimation Method

A cognitive task was employed in this study in which a task worker performs repeatedly problems which can be solved in several seconds. The task performance was measured by counting the number of solved problems as well as the above physiological feature value extractions. Then task performance was estimated by regression analysis with using physiological features (i.e. LF, HF, LF/HF, and MPD) and they were compared with the measured task performance. There are various methods of regression analysis and multiple regression analysis is one of the famous methods. However, it is assumed that regression line cannot be calculated accurately by the analysis because of multicollinearity. On the other hand, there are various studies where machine learning methods are employed as the physiological analysis [20-25]. In this study, therefore, SVR and RF were employed because these methods are supposed to avoid multicollinearity problem and have high generalization capability.

### 3.1 Support Vector Regression (SVR)

SVR is one of the non-linear regression analysis methods based on Support Vector Machine (SVM), which is effective for 2-class classification and was employed so as to estimate emotions by using several physiological indices [21,22] and to detect stress [23].

The method to generate SVM classifier is described below. If  $n$  features  $\mathbf{x}_i (i = 1, \dots, n)$  belong to class  $y_i \in \{-1, 1\}$ , in order to classify them correctly, the hyperplane  $\mathbf{w}^T \mathbf{x} + b = 0$  ( $\mathbf{w}^T$  is a coefficient vector,  $b$  is a bias term) is calculated which maximizes the distance between the hyperplane and the support vector  $\mathbf{x}^*$  which is the

closest to the hyperplane of all  $\mathbf{x}_i$ . The SVM model (the above hyperplane) can be calculated by solving this problem. The SVR model can be also calculated by using this hyperplane and solving regression problem.

### 3.2 Random Forest (RF)

RF is the machine learning algorithm proposed by Breiman [26] and is used to solve classification or regression problems as well as SVM and SVR. RF is the application method of the classification method called “bagging”, which generates a number of weak learners by using a part of training data and integrate these weak learners by the majority rule. When estimating cognitive states by several physiological indices, it is said that RF and AdaBoost, which is similar to RF, are as accurate as SVM [24,25].

The method to generate RF regression model is described below. First,  $n$  data is extracted with allowing duplication from  $n$  training data  $\mathbf{x}_i (i = 1, \dots, n)$  ( $\dim \mathbf{x}_i = d$ ) and these  $n$  data is defined as new training data (bootstrap sampling). Next, a decision tree is generated by using this training data, and  $d'$  ( $< d$ ) features are selected randomly in each non-terminal node of the tree. If several decision trees are generated as above, these trees have a weak correlation. After generating  $m$  decision trees, RF training finishes. The majority or the average of outputs from these trees is defined as the estimation result in the case of classification or regression problem solving, respectively. In training phase, the decision of the size of tree node,  $d'$ , and  $m$  are required before training. It is known that the node size and  $d'$  are recommended to be 5 and  $d/3$ , respectively in the case of regression problem [26]. On the other hand,  $m$  may be large enough because overlearning is not caused even though  $m$  is too large. In this study, these values were employed when employing RF as the estimation method.

### 3.3 Model Calculation

In training phase, SVR and RF models were calculated by using physiological data of one task worker in order to deal with individual difference of physiological responses. Moreover, MATLAB®[27] was used to calculate models and LIBSVM and random-forest-matlab were used as machine learning library.

## 4 Experiment

### 4.1 Purpose

The purpose of this experiment was to validate heart rate variability and pupil diameter as explanatory variables of task performance estimation and to evaluate the accuracy of the proposed methods.

## 4.2 Participants

31 healthy volunteers participated in this experiment and their native language was Japanese. In this experiment, only those who do not wear glasses participated because the accuracy of eye tracking system gets lower when wearing glasses.

## 4.3 Measurement of Physiological Indices

In the case of measurement of heart rate, ECG electrodes were pasted on a left rib and a right clavicle. Ground and reference electrodes were pasted on right and left earlobes, respectively. In order to reduce noise signals, the time constant of high-pass filter and the cut-off frequency of low-pass filter were set to 3.0 sec and 100Hz, respectively. Furthermore, notch filter was set to 60Hz so as to reduce hum noise from commercial power supply.

Pupil diameter was measured by an infrared eye tracking camera, faceLAB5 [12]. A participant sat on the chair while performing tasks on the desk. Then, the position and height of the chair were adjusted and the angle of the eye tracking camera was set to 36 degrees against the desk so as to recognize the face of a participant correctly.

## 4.4 Cognitive Task and Instruction of Task Performing

In this study, Receipt Classification Task [28] was used as a cognitive task to evaluate task performance. The task was presented on PC display.

In order to avoid overlearning and generate regression models properly, various values of training data should be measured and provided to machine learnings as training data. It is therefore necessary to instruct participants to change task performance (i.e. the number of classified receipts) and collect their physiological indices while their task performance varies. In this study, the participants were instructed to change the speed of classifying receipts according to the indicator of color bar displayed on the left as shown in Fig. 1. Concretely, the slower they were instructed to change the classifying speed, the larger the ratio of blue in the color bar got. On the contrary, the faster they changed the speed, the larger the ratio of red got. At the beginning of the task, the color bar was red or blue. Then, the ratio of the other color got larger and larger at a constant rate. Finally, the ratio got 100% 30 minutes after the beginning. In this study, the case in which the color changes blue into red was called "Pace-up" Phase, while the opposite case was called "Pace-down" Phase.

## 4.5 Experimental Protocol

The experiment was conducted for 2 days. The purpose of the first day experiment was (1) to perform participant screening, (2) to get accustomed to experimental environment, and (3) to practice Receipt Classification Task and Pace-up and Pace-down Phase. Regarding (1), the accuracy of eye tracking could be low because the camera could not capture eyes of some participants enough if eyes of them were small. In this case, they don't participate in the second day experiment. In regard to (2), it was supposed that participants got stressed because they participated in the experiment and performed the cognitive task for the first time, which may affect their physiological responses. In order

to reduce this psychological influence, the first day experiment was designed to make them adapt to this environment. Finally, regarding (3), if participants perform the cognitive task for the first time, cognitive load on them may be high even if the difficulty of the task was low. However, if they get used to the task, cognitive load may be lower even though the difficulty was the same. This is generally called “Learning Effect”. This effect should be reduced because it gives an influence on physiological indices. In order to remove this effect, they practiced the task, Pace-up, and Pace-down Phase in the first day experiment.

**Table 1.** Experimental schedule: second day

Duration (minutes)	Content
10	Setting of Electrodes and Eye Tracking Camera
10	Task Practice
30	Phase A*
10	Rest
30	Phase B*
5	Removal of the Instruments

\*: Measurement of Physiological Indices and Task log.

The experimental schedule of the second day is shown in Table 1. Phase A in this table was set to either Pace-up or Pace-down Phase at random while Phase B was set to the other in order to get counterbalance of ordering effect. Physiological indices and task log were recorded in these phases.

## 5 Result

### 5.1 Participant Screening

As the result of screening in the first day experiment, pupil diameters of all the participants could be measured correctly by the eye tracking system. Then all of them participated in the second day experiment. However, the heart rate variability data of 4 participants were not properly measured because strong artifacts appeared. Thus, the experimental results of 27 participants were analyzed, where 13 participants performed Pace-up Phase in Phase A, while the other 14 did it in Phase B.

### 5.2 Task Performance Estimation

After the experiment, SVR and RF models which estimate task performance (i.e. the number of classified receipts) were calculated every participant by using the physiological features such as LF, HF, LF/HF and MPD. Then, the accuracies of these models were evaluated by using 13-fold cross validation.

Average of coefficients of determination  $R^2$  of each method are described in Table 2.

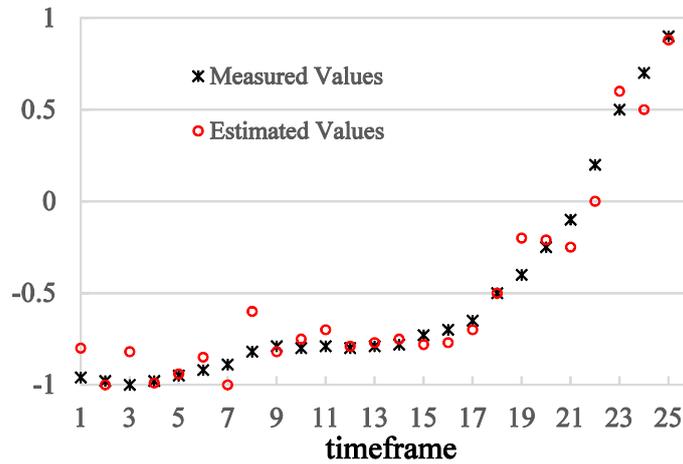
As the result of paired and two-tailed t-test,  $R^2$  of the SVR model was significantly higher than that of RF ( $p < 0.001$ ).

**Table 2.** Average of  $R^2$  of each method.

Method	$R^2$	
	Average	SD
SVR	0.875*	0.097
RF	0.648	0.199

\*:  $R^2$  of SVR was significantly higher than that of RF ( $p < 0.001$ )

An example of the SVR estimation is shown as Fig.1 ( $R^2 = 0.969$ ). The values of vertical axis in Fig. 1 was normalized to [-1,1]. “Measured Values” in Fig.1 is the number of classified receipts and “Estimated Values” is the one estimated by SVR.



**Fig. 1.** An example of the SVR estimation (Pace-up Phase).

## 6 Discussion

### 6.1 Estimation Methods

As the comparison of the result between the SVR and RF models, the SVR model has significantly better estimation performance in terms of  $R^2$ . RF is the ensemble learning method which reduces generalization error by suppressing the variance with keeping the bias of model low. In order to achieve it, RF generates a number of decision trees which have weak relationship to each other by using bootstrap sampling. However, if the number of training data is small, RF cannot generate various decision trees enough to estimate response variables accurately. In this study, there were 52 feature data per a participant. Finally, only 48 training data per a participant remained because 13-fold

cross validation was employed ( $52 \times 12 / 13 = 48$ ). Regarding RF, the number of this sample, 48, was not enough to estimate task performance accurately. On the other hand, in the case of the SVR model, the number of tunable parameters is more than the RF model. Therefore, the performance of the SVR model is supposed to be significantly more accurate in this study.

## 6.2 Physiological Indices

In this study, the contribution of feature values (LF, HF, LF/HF and MPD) was considered in the case of the SVR model. These contributions can be discussed by evaluating each component of the coefficient vector  $\mathbf{w}$  in the hyperplane  $\mathbf{w}^T \mathbf{x} + b = 0$  calculated by SVR training. Average coefficients of the feature values are shown in Table 3. It was found that the contribution of MPD was the highest among the feature values and that of LF was the second highest.

**Table 3.** Average coefficients of feature values.

	MPD	LF	LF/HF	HF
Average of Coefficients	2.00	-0.98	-0.71	-0.27

Regarding the sign of each coefficient, MPD had positive relation to task performance. The result was supported by the study conducted by Poock [18] which shows the positive relationship between the size of pupil diameter and the speed of information processing. In the case of Receipt Classification Task, when the speed of performing task got faster, it is assumed that it requires parallel information processing in which participants should memorize several components (e.g. the date, amount, and name of company of a receipt). Then, it was supposed that their pupil diameter got larger [6,7] because of the high cognitive load. On the other hand, LF had negative relation to task performance. According to Mulder [16], the higher the difficulty of a cognitive task gets, the lower the power of LF gets. This supports the result of this study.

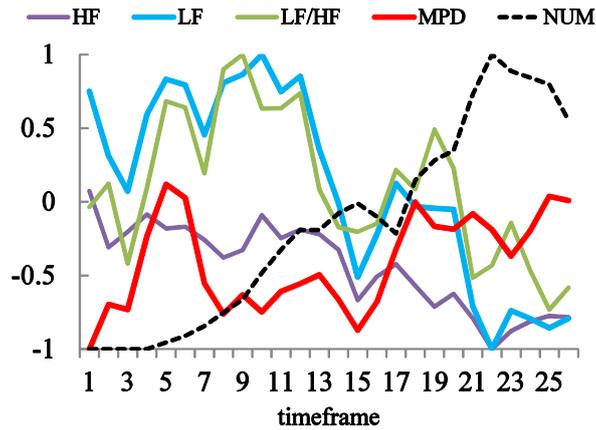
## 6.3 Individual Difference

As mentioned above, the contribution of MPD was the highest, while MPD of some subjects had a weak correlation with task performance and heart rate variability of them had a strong correlation. Table 4, Fig.2, and Fig.3 show an example (EX1). The values of vertical axis in Fig. 2 and Fig.3 were normalized to [-1,1].

Comparing Table 4 with Table 3, it shows coefficients of EX1 are different from the averages by individual difference. Therefore, if the estimation method such as SVR and RF employs a single explanatory value, the estimation performance can be low. In the case of the SVR model which employed these feature values,  $R^2$  in EX1 was 0.878, respectively. The performance is so high, which suggested that plural physiological indices make task performance estimation accurate even if the individual differences are found.

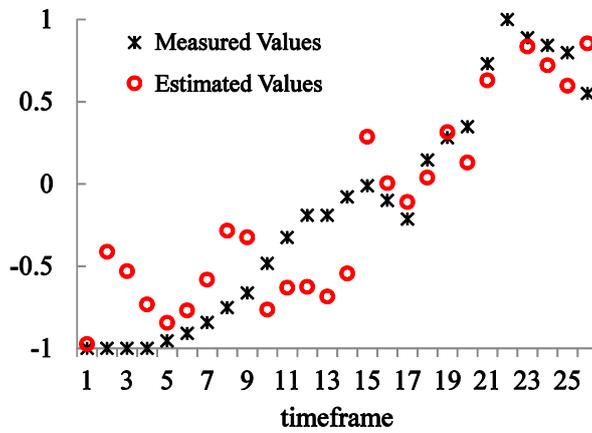
**Table 4.** Coefficients of feature values in EX1.

	MPD	LF	LF/HF	HF
Coefficients	-0.02	-3.05	-2.06	-3.38



※NUM = the number of classified receipts

**Fig. 2.** Physiological features and task performance in EX1.



**Fig. 3.** An example of the SVR estimation in EX1 (Pace-up Phase).

## 7 Conclusion

In this study, the method which estimates task performance of intellectual workers by using several physiological indices (pupil diameter and heart rate variability) was proposed in order to develop the objective and quantitative evaluation method of the intellectual productivity. As the physiological indices for task performance estimation, heart

rate variability and pupil diameter were employed, which were supposed to reflect human cognitive states (e.g. stressed and relaxed). These indices are also expected to be measured by contactless measurement method without giving any load. As the estimation model of task performance, two machine learning models, Support Vector Regression (SVR) and Random Forests (RF), have been employed. As the result of a subject experiment, it was found that coefficient of determination ( $R^2$ ) of SVR was 0.875 which was higher than that of RF ( $p < 0.01$ ). The result suggested that pupil diameter and heart rate variability were effective as the explanatory variables and SVR estimation was also effective in task performance evaluation.

However, the cognitive task, Receipt Classification Task, which was employed in this study, requires only a part of the cognitive processing used by office works. It is therefore necessary to consider the accuracy of the proposed method with various cognitive tasks. In the future, the authors are aiming at developing the more accurate evaluation method of the intellectual productivity.

## Acknowledgement

This study was supported by JSPS KAKENHI Grant Number 23360257.

## References

1. Indoor Health and Productivity Project, Lawrence Berkeley National Laboratory, <https://iaqscience.lbl.gov/>, Accessed Jan. 2017.
2. R. Brünken, J. L. Plass, and D. Leutner: Direct Measurement of Cognitive Load in Multimedia Learning. *EDUCATIONAL PSYCHOLOGIST*, Vol. 38, No. 1, pp. 53-61 (2003)
3. F. Paas, J. E. Tuovinen, H. P. Tabbersm, and W. M. Gerven: Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *EDUCATIONAL PSYCHOLOGIST*, Vol. 38, No. 1, pp. 63-71 (2003)
4. F. Paas, and J. J. G. Merriënboer: The efficiency of instructional conditions, An approach to combine mental effort and performance measures. *Human Factors*, Vol. 35, No. 4, pp. 737-743 (1993)
5. P. Ayres, and F. Paas: Cognitive Load Theory, New Directions and Challenges. *Applied Cognitive Psychology*, Vol. 26, No. 6, pp. 827-832 (2012)
6. E. H. Hess, and J. M. Polt: Pupil size in relation to mental activity during simple problem-solving. *Science*, Vol. 143, pp. 1190-1192 (1964)
7. H. Yamamoto, and Y. Shimizu: Change of Pupil Size and Blinking Time in Memorizing Numerals. *The Journal of the Institute of Television Engineers of Japan*, Vol. 45, No. 3, pp. 370-376 (1991)
8. T. Ishibashi, A. Ohtani, and T. Miura: Heart Rate as an Index of the Mental Load. *Japan Society for Occupational Health*, Vol. 10, No. 7, pp. 377-379 (1968)
9. D. O. Kennedy, A. B. Scholey: Glucose administration, heart rate and cognitive performance, effects of increasing mental effort. *Psycho-pharmacology*, Vol. 149, No. 1, pp. 63-71 (2000)
10. N. Nishihara, P. Wargocki, and S. Tanabe: Cerebral blood flow, fatigue, mental effort, and task performance in offices with two different pollution loads. *Building and Environment*, Vol. 71, pp. 153-164 (2014)

11. S. Kunimasa, K. Miyagi, H. Shimoda, and H. Ishii: Detection of Temporary Rest State while Performing Mental Works by Measuring Physiological Indices. The transactions of Human Interface Society, Vol.16, No.1, pp. 41--50. (2014)
12. facelab, Seeing Machines, Inc., <http://www.seeingmachines.com/>, Accessed Jan. (2017)
13. M. Z. Poh, D. J. McDuff, and R. W. Picard: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics Express, Vol. 18, No.10, pp. 10762--10774 (2010)
14. P. G. A. M. Jorna: Spectral analysis of heart rate and psychological state, A review of its validity as a workload index. Biological Psychology, Vol. 34, No. 2, pp.237--257 (1992)
15. U. R. Acharya, P. K. Joseph, N. Kannathal, C. Min Lim, and J. S. Suri: Heart rate variability, a review. Medical and Biological Engineering and Computing, Vol. 44, No. 12, pp. 1031--1051 (2006)
16. G. Mulder, and L. J. M. Mulder: Information Processing and Cardio vascular Control. Psychophysiology, Vol. 18, No. 4, pp. 392--402 (1981)
17. S. Saida, M. Akamatsu, Y. Inukai, Y. Kuchinomachi, K. Nakamura, N. Nagamura, and M. Yoshioka: Human Measurement Handbook. Asakura Publishing Co., Ltd. (2003)
18. G. K. Pooek: Information processing vs pupil diameter. Perceptual and Motor Skills, Vol. 37, No. 3, pp. 1000--1002 (1973)
19. D. Kahnemann, and J. Beatty: Pupillary responses in a pitch-discrimination task. Perception and Psychophysics, Vol. 2, No.3, pp. 101--105 (1967)
20. R. W. Picard, E. Vyzas, and J. Healey: Toward machine emotional intelligence, Analysis of active physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No.10, pp. 1175--1191 (2001)
21. K. Takahashi: Remarks on SVM-based emotion recognition from multi modal biopotential signals. 13th IEEE International Workshop on Robot and Human Interactive Communication, pp. 95--100 (2004)
22. C. D. Katsis, N. Katertsidis, and G. Ganiatsas: Toward Emotion Recognition in Car-Racing Drivers. A Biosignal Processing Approach. IEEE Transactions on Systems, Man and Cybernetics-Part A, Systems and Humans, Vol. 38, No.3, pp. 502--512 (2008)
23. J. Zhai, and A. Barreto: Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables. Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE, pp. 1355--1358 (2006)
24. C. Zhao, C. Zheng, M. Zhao, and J. Liu: Physiological Assessment of Driving Mental Fatigue Using Wavelet Packet Energy and Random Forests. American Journal of Biomedical Sciences, Vol. 2, No.3, pp. 262--274 (2010)
25. M. Miyaji, H. Kawanaka, and K. Oguri: Driver's Cognitive Distraction Detection Using Physiological Features by the AdaBoost. Proceeding of the 12th International IEEE Conference on Intelligent Transportation Systems, pp. 90--95 (2009)
26. L. Breiman: Random forests. Machine learning, Vol. 45, No.1, pp. 5--32 (2001)
27. Inc.MathWorks,MATLAB. <https://www.mathworks.com/products/matlab.html>, Accessed Jan. (2017)
28. K. Miyagi, K. Oishi, K. Uchiyama, H. Ishii, H. Shimoda: Proposal of intellectual productivity model based on work state transition. Conference on HumanComputer Interaction, Engineering Psychology and Cognitive Ergonomics, 15th International Conference on Human-Computer Interaction, Vol.16, pp.335--343 (2013)