This advance publication article will be replaced by the finalized version after proofreading.
Measuring Mental Workload of Software Developers Based on Nasal Skin Temperature

Keitaro NAKASAI†, Shin KOMEDA††, Nonmember, Masateru TSUNODA†††, and Masayuki KASHIMA††††, Member

SUMMARY To automatically measure the mental workload of developers, existing studies have used biometric measures such as brain waves and the heart rate. However, developers are often required to equip certain devices when measuring them, and can therefore be physically burdened. In this study, we evaluated the feasibility of non-contact biometric measures based on the nasal skin temperature (NST). In the experiment, the proposed biometric measures were more accurate than non-biometric measures.

Key words: Mental workload, NASA-TLX, affective engineering

1. Introduction

Using biometric measures, some studies have recently attempted to measure the mental workload of developers during software development [3][8]. Mental workload refers to the extent of mental burden of a person engaged in a task. When the mental workload is high, it suggests that the person is coping with mentally difficult tasks.

A continual high mental workload may cause human errors [9]. If we identify such a high mental workload of developers, we could decrease the risk of errors by warning developers of such a workload. However, it is not easy for other people to recognize the mental workload of a developer based on appearance. Hence, biometric measures based on brain waves and heart rate are often used to measure the mental workload of developers [3][8].

To conduct such measurements, developers are often required to equip certain devices such as a headgear. They may experience a certain amount of physical burden to equip such devices, and some of them would not want to do so. Hence, non-contact biometric measures are needed, when we introduce that into broader development projects.

Therefore, we focused on biometric measures based on nasal skin temperature (NST). NST of an individual can be measured using thermography. When thermography is used, developers are not required to install any device. NST has been widely used in other fields to measure mental status [1][7].

While programming consists of the following phases, unlike the mental arithmetic applied in [7] and reading contents performed in [1]:

- Planning the structure of the source code.
- Inputting the source code based on the plan.
- Removing defects included in the code.

The extent of the mental workload in each phase may differ. To consider such programming activities, we proposed new NST-based metrics. The goal of our study is to distinguish the difficulty of the task with which developers cope. To achieve that, we proposed and evaluated the NST metrics. Thus, the proposed NST metrics and evaluation are the major contributions of our study.

2. Nasal Skin Temperature

Overview: When the sympathetic nervous system is active (i.e., the mental status of a person is active), blood flow and skin temperature decrease. There are peripheral blood vessels just under the nose, and NST is significantly affected by blood flow. Hence, NST can indirectly observe the activity of the sympathetic nervous system [1]. When NST is low, the mental status of a person is active. The measurement target of NST is same as heart rate variability (HRV) measured by electrocardiogram (ECG), but NST can avoid drawbacks of ECG such as influences of artifacts.

NST has been widely used in other fields to measure a mental status [1][7]. Previous studies have used this approach to measure the mental workload. To continuously increase their mental workload, Mizuno et al. [7] proposed assigning a 3-min mental arithmetic task to their participants six times. Based on past studies in other fields, NST is expected to be useful in measuring the mental workload during software development.

However, as explained in Section 1, programming comprises various phases, and the extent of the mental workload in each phase might be different. Hence, it is unclear how NST is effective in measuring the mental workload during programming. Unlike past studies, one of the major contributions of our study is to show the effectiveness of NST empirically to measure the mental workload of developers in programming. Based on studies [2][5], although many studies tried to measure mental workload, NST has not been used.

The other major contribution is to propose NST metrics, compared with past studies such as [1] that estimate cognitive load using NST. While the task time in study [1] was 3 minutes, ours was 13 minutes on average, and the median value was 11 minutes (shown in Table 2). That is,
the task time is three times larger than study [1] at least, and hence, we defined NST metrics considering the variance.

Comparison to other biometric sensors: Table 1 summarizes influence of artifacts and extent of burden to participants on major biometric sensors, which were selected based on study [5]. We consider that the extent of burden on electroencephalogram (EEG) and ECG is large, because a headgear is needed. The extent on functional magnetic resonance imaging (FMRI) is very large, because participants lie in the device, and they cannot move during the measurement. The influence of artifacts on EEG is large [6]. We regard the influence on ECG and FMRI as medium, because although they are affected by artifacts such as muscular ones, they can be handled to some extent.

We set the influence of artifacts on photoplethysmography (PPG) as large, since HRV based on PPG is inaccurate [4]. While smartwatch is often used as PPG device, and therefore the extent of burden is medium. To use eye-tracker and thermography, participants are not required to equip any devices. Hence, the extent of burden is low on them. At least, they are not affected by movement of participants much. Therefore, we think influence of artifacts is small on them.

Metrics for Estimation of Cognitive Load: Because the forehead skin temperature is not significantly affected by blood flow, and the difference between the forehead and NST signifies the status of the sympathetic nervous system. We calculated the difference to use NST as a biometric measure. It has frequently been applied in previous studies such as [1]. Although one study [1] named this the nose-forehead difference, we call denote as DIFF, to shorten the term.

Using DIFF, we defined the following NST metrics:

- **GAP**: DIFF during a task minus DIFF while at rest
- **WMAX**: Maximum value of GAP
- **WAVE**: Average value of GAP
- **WSUM**: Sum of GAP

In our experiment, DIFF differed among the participants. To normalize that, we treated DIFF while at rest as the baseline, and calculated the gap from the baseline, same as study [1]. We call it GAP. As described in Section 2, the programming consists of different phases, and the mental workload may differ among phases. Hence, we assumed that the GAP varies during a task, and WMAX is defined to focus on the maximum workload on the task. WAVE is defined to consider the average mental workload during a task. WSUM considers the total mental workload during the task.

Figure 1 shows an example of these metrics. For example, the forehead and nasal-skin temperature were recorded every 2 min, and the metrics were calculated at each point in time. In the example, we assumed that the DIFF at rest was 1.5.

3. Experiment

Tasks: In the experiment, the participants engaged in the following three programming tasks:

- **Preliminary**: (printing “hello world”) to learn the experiment environment.
- **Easy**: (printing a multiplication table using a for loop).
- **Difficult**: (counting the digits of the given numbers).

We selected the tasks from Aizu Online Judge (https://judge.u-aizu.ac.jp/onlinejudge/). The order of easy and difficult tasks was randomly changed for each participant. The participants rested for 3 min before each task to drop the NST. The rest time was the same as that in study [7].

Procedure: The participants edited the programs on a web browser. The program was tested using JUnit. When they clicked the execute button, the execution result of the program and the output of JUnit were shown in the browser. We recorded each task time (i.e., from the time to show the specifications of the program to the time to pass the JUnit test). After each task, the participants answered a questionnaire.

Questionnaire: As the questionnaire to measure the mental workload of participants, we used simplified NASA-TLX [12]. Same as NASA-TLX, the questionnaire measures the mental demand, physical demand, temporal demand, own performance, effort, and frustration. It measures not only the mental workload but also the physical workload. Hence, past studies such as that described in [13] used some of the metrics. Similarly, we used mental demand, own performance, effort, and frustration level. We call them workload scores. The workload scores are based on the average of 7-point scales. They ranged from -3 to 3, wherein lower values signify higher workload of participants.

Participants: Participants were seven undergraduate (fourth year) and master course students majoring in computer science. They have learned Java language over three years, and also studied design patterns. Note that the number of data points was 14, since each participant coped with two tasks. While presenting the dataset in a table, the variables are included in the columns whereas the data points are presented in the rows.

Experimental device: We used a FLIR T530 for thermography. The distance between the thermograph and the participants was approximately 1 m, and the emissivity was set to 0.98. Although cost of the devices was high, there
are low cost devices. Because we assume that the programming phases described in Section 2 are not frequently switched to other phases, and the NST gradually changes, we calculated NST metrics every 2 min. During the experiment, the head of each participant was not fixed by chin-rest. However, we picked up NST only when participants directed their face to display (i.e., thermography). Hence, the angle of the face did not affect the result.

4. Results

Discrimination of tasks: We tried to distinguish the difficulty of the task, based on NST metrics. We stratified the experimental results into easy and difficult tasks. After the stratification, we calculated the average and the median of time, NST metrics, and workload scores on each task. We focused on the difference of the average and the median between the tasks.

On the results, average GAP was 0.70, and the maximum and minimum values were 1.3 and 0.26 respectively. Based on GAP, NST metrics were calculated. Table 2 and 3 show the summary. In the tables, bold face denotes the number is smaller than the other task. Most of workload scores on difficult task were smaller than easy one. Especially, the difference of the median was almost 1.0 on performance and frustration level. When the mental workload is higher, the scores are smaller. Therefore, “difficult” task was actually difficult for the participants, based on the scores. Also, time of the difficult task was longer than the easy one. That also suggests the difficult task was actually difficult.

When we focus on NST metrics, they were larger when the task was difficult. When the mental status is active, NST metrics becomes high. That is, the result suggests that NST metrics reflect the difficulty of the tasks properly. Hence, we can distinguish the difficulty of the task based on the metrics.

Prediction methods: To evaluate prediction performance of NST metrics to workload scores, we made the following methods:

- **NST**: Linear regression on which candidates of independent variables are task time, WMAX, WAVE, and WSUM.
- **Seconds**: Linear regression on which independent variable is task time only.
- **Mean**: Average of each workload score.
- **Q2**: Median of each workload score.

Mental demand, own effort, effort, and frustration level measured were set as the dependent variables of each method. Seconds, mean, and Q2 methods are based on non-biometric measures (i.e., existing approaches).

Mean method used the average of mental demand as a predicted value, for instance. The average was calculated based on the leave-one-out cross validation. Q2 (i.e., median) method was applied in the same manner. To prepare seconds and NST method, we applied a log transformation to the task time. On NST method, we applied a stepwise variable selection. To consider multicollinearity, when the tolerance was smaller than 1.0, the variable was not added to the model during selection.

Four models were developed for each prediction method.

The candidates for the independent variables were the same, but those for the dependent variables were different. For example, we created four models using the NST method, wherein each model predicted mental demand, performance, effort, and frustration levels.

Evaluation criteria: To evaluate the methods, we used the average and the median of absolute errors between actual and predicted values. To derive prediction values of each method, we applied leave-one-out cross validation.

We also compared seconds and NST method using adjusted $R^2$. It evaluates the explanatory power. The power increases according to the number of independent variables. Adjusted $R^2$ is used to avoid the influence of the number. When the value is greater than 0.5, the model explains the dependent variables well.

Result: Table 4 shows the average and the median of absolute error. In the table, bold faces signify that the error was the smallest among the methods. As shown in the table, the average and the median of the NST method were the smallest on the most cases. That is, NST method was more effective than non-biometric measures for the prediction.

Table 5 shows the adjusted $R^2$ values for seconds and NST method. All adjusted $R^2$ of NST method were higher than seconds method. However, adjusted $R^2$ was very small when mental demand was predicted. Additionally, they were statistically significant at the 0.05 level except for the mental demand. Except for that, NST metrics had explanatory power for workload scores to some extent.

Table 6 shows the standardized partial regression coefficients of each model. In the table, “.” indicates that the variable was not included in the model. Although some of the coefficients of NST metrics were not statistically significant at the 0.05 level, the absolute value of the coefficients was larger than one-half of the coefficients for time. The results suggest that the relationships of NST metrics to mental workload are not insignificant, compared with those for time. While signs of some partial regression coefficients were positive. That is, they did not follow the assumption that time and NST metrics increase when workload scores decrease (i.e., mental workload increases). Although we performed variable selection considering multicollinearity, it might still remain. This suggests that there is room for improvement, when we predict the mental

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Average and median of time and NST metrics on each task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>Time (seconds)</td>
</tr>
<tr>
<td>Average</td>
<td>Easy</td>
</tr>
<tr>
<td></td>
<td>Difficult</td>
</tr>
<tr>
<td>Median</td>
<td>Easy</td>
</tr>
<tr>
<td></td>
<td>Difficult</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Average and median of workload scores on each task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>Mental demand</td>
</tr>
<tr>
<td>Average</td>
<td>Easy</td>
</tr>
<tr>
<td></td>
<td>Difficult</td>
</tr>
<tr>
<td>Median</td>
<td>Easy</td>
</tr>
<tr>
<td></td>
<td>Difficult</td>
</tr>
</tbody>
</table>
workload of developers by NST metrics.

The findings of our experimental results are summarized as follows:

- We could distinguish the difficulty of the task, based on NST metrics (based on Table 6).
- The accuracy of NST metrics is expected to be better than non-biometric measures, when predicting the mental workload of developers (based on Table 5).

5. Discussion

**Threats to Validity:** The participants in the experiment were not professional software developers. Previous studies have shown that students instead of professionals can be used in such experiments [11]. Therefore, we consider that the results of our study would not have been extremely different, when professionals participated in our experiment. The recruitment of professional developers will be an area of future study, however.

In our experiment, we used only a small number of specifications owing to time limitations. However, it is difficult to use realistic programs in such experiments. For instance, one study [10] used codes whose lines ranged from 17 to 32. However, the size must be considered when interpreting the results.

The number of participants was small. However, it is not easy to increase the number in this type of study. For example, 10 participants were included in the study [7][10]. We acknowledge that the number of participants should be increased to enhance the reliability of the results. While NST has been widely used in other fields to measure a mental status, as explained in Section 2. Therefore, we believe that measuring mental workload of developers by NST is promising, and the results have external validity to some extent.

**Use case:** As suggested in study [2], when the mental workload of a developer is high, the difficulty of coding is high. As a result, defects could be injected into the task with high probability. If we automatically measure the mental workload of developers, we can identify such defect-prone tasks automatically.

Based on the identification, we do not review all tasks equability, but defect-prone tasks are automatically identified and thoroughly reviewed. As a result, the review efficiency (i.e., productivity) and probability of finding defects (i.e., quality) are expected to be high. The advantages are mentioned in study [2]. That is, our approach can be applied to coding, enhancing productivity and quality of code review.

Considering the size of the source code in the experiment, if we regard the coding of each method as a task, our approach can be applied to the task. When we automatically record coding history and NST metrics with timestamps, we can relate NST metrics to the coding for each method. As a result, the mental workload of each task can be measured.

6. Conclusion

We propose to use the nasal skin temperature (NST) to measure mental workload of developers, as non-contact biometric measures. For the measurement, we defined new metrics WMAX, WAVE, WSUM based on NST. To broaden the measurement of mental workload of developers, non-contact measures are needed. To evaluate the feasibility of NST, we conducted an experiment in which the participants engaged in easy and difficult programming tasks.

The results suggest that the difficulty of a task can be assessed based on the mental workload predictions using the NST metrics. Additionally, their prediction accuracy is expected to be better than that of the non-biometric metrics.

As our future studies, we will evaluate the combination of NST metrics with other non-contact biometric measures (e.g., eye-gaze) to predict mental workload of developers.

References


